

KNOWING WHY

Essays on Self-Knowledge and Rational Expalnation

Ryan Cox

A thesis presented for the degree of
Doctor of Philosophy



Department of Philosophy
Research School of Social Sciences
Australian National University
Australia

ABSTRACT

This thesis is about basic self-knowledge, our knowledge of particular facts about our own actions, attitudes, and sensory states. It focuses on our knowledge of particular explanatory facts about our own actions and attitudes and defends a particular interrogative theory of such knowledge, according to which our knowledge of such facts about our own actions and attitudes is arrived at by resolving related practical questions about our own actions and attitudes. It consists of four related essays.

Essay one, "Introspection Without Inferentialism", argues against the view that a belief is rational only if it is based on sufficient reasons for so believing. It argues that beliefs about our own attitudes and explanations thereof can be rational even though they are not based on sufficient reasons for so believing. This conclusion clears the way for alternative theories of how we arrive at rational beliefs about our own actions and attitudes, like the theory to be defended in the third and fourth essay.

Essay two, "Actions, Reasons, and Because", is about the relation between two different kinds of explanations of our own actions and attitudes: reason explanations and psychological explanations. It argues on broadly linguistic grounds that reason explanations are a kind of psychological explanation and are therefore not a distinct kind of rational explanation. A consequence of this conclusion is that questions about how we arrive at knowledge of particular reason explanations can be treated as a special case of how we arrive at knowledge of psychological or rational explanations more generally.

Essay three, "Knowing Why", is about our knowledge of particular explanatory facts about our own actions and attitudes. It argues against the view that our knowledge of particular rational explanations of our own actions and attitudes is inferential and defends an alternative theory according to which such knowledge is non-inferential. According to the alternative theory, our knowledge of particular rational explanations of our own actions and attitudes is arrived at by resolving related practical questions about our own actions and attitudes. I argue that the latter theory provides a better explanation than an inferential theory of observed self and other symmetries concerning knowledge of particular rational explanations of our actions and attitudes.

Essay four, "Interrogative Theories of Introspection", develops the theory defended in essay three in more detail and argues that it offers a plausible general theory of how we arrive at knowledge of particular facts about our own actions and attitudes. It explicates three distinct interrogative theories of introspection and argues that a theory on which we can arrive at knowledge of particular facts about our actions and attitudes by resolving related practical questions about those actions and attitudes provides the the best explanation of how we can come to know particular facts about our actions and attitudes in a non-inferential way.

DEDICATION

For my mum and dad.

DECLARATION

I declare that what follows is entirely my own work.

ACKNOWLEDGEMENTS

I would like especially to thank Daniel Stoljar for his supervision. I couldn't have asked for a better supervisor. I have learnt most of what I know about how to do philosophy from Daniel. I'd also like to thank the other members of my supervisory panel, David Chalmers, Daniel Nolan, and Nicholas Southwood. I'd like to thank Michael Smith for his supervision while I was in Princeton over the Fall of 2012 and Spring of 2013 terms and I'd like to thank John Maier for his mentorship while I was visiting Cambridge in 2014. For comments, feedback, and discussion, I'd like to thank, Daniel Stoljar, David Chalmers, Daniel Nolan, Nicholas Southwood, Frank Jackson, John Maier, Leon Leontyev, Johannes Himmelreich, Derek Baker, Gideon Rosen, Daniel Wodak, Mark Harris, David Velleman, Naomi Kloosterboer, Richard Holton, Rachael Briggs, Luke Roelofs, Erick Llamas, Daniel Gregory, Luara Ferracioli, Holly Lawford-Smith, Matthew Hammerton, Christopher Gyngell, Alexander Sandgren, Lucy Campbell, Edward Elliott, Eleni Saava, Richard Holton, and Jane Heal. Finally, I'd like to thank Luara Ferracioli for her constant support and love without which I couldn't have done this, and I'd like to thank Valentina Ferracioli Cox for making it all worthwhile.

CONTENTS

1	Introduction	1
1.1	The Problem of Self-Knowledge	3
1.1.1	Reasons and Rationality	4
1.1.2	Knowledge of Rational Explanations	6
1.1.3	Interrogative Theories of Introspection	8
2	Introspection Without Inferentialism	10
2.1	Introduction	10
2.2	Inferentialism	13
2.2.1	Factual and Psychological Basing	15
2.2.2	Conservative, Moderate, and Radical Inferentialism	16
2.3	The Case Against Moderate Inferentialism	18
2.4	The Case Against Radical Inferentialism	25
2.4.1	Cassam's Radical Inferentialism	25
2.4.2	Zimmerman's Radical Inferentialism	27
2.5	Conclusion	30
3	Actions, Reasons, and Because	32
3.1	Introduction	32
3.2	The Language of Reason Explanations	34
3.2.1	Explanation and Explanatory Language	36
3.2.2	Overview	37
3.3	The Conversational Implicature Hypothesis	39
3.3.1	Cancelability and Reinforceability	42
3.3.2	Calculability	45
3.3.3	Embedded Contexts	49
3.4	The Pragmatic Presupposition Hypothesis	50
3.5	The Conventional Implicature Hypothesis	54
3.6	The Semantic/Pragmatic Underdetermination Hypothesis	57

3.7	The Syntactic Ambiguity Hypothesis	58
3.8	The Lexical Ambiguity Hypothesis	60
3.8.1	Because Clauses as Subject Oriented Adverbials	61
3.8.2	Linguistic Tests for Ambiguity	64
3.8.3	Multiplying Senses Beyond Necessity	66
3.9	The Context-Dependence Hypothesis	67
3.10	Conclusion	68
4	Knowing Why	70
4.1	Introduction	70
4.2	The Argument from Symmetry	73
4.2.1	The Evidence for Symmetry	81
4.2.2	Problems With the Evidence for Symmetry	85
4.3	Against Inference	89
4.4	The Tension	92
4.5	An Alternative Explanation	93
4.5.1	The Practical Question Theory	94
4.5.2	Extending the Practical Question Theory	95
4.5.3	Simulation and the Practical Question Theory	99
4.5.4	Explaining Symmetry	99
4.6	Conclusion	102
4.7	Appendix: Further Evidence	103
4.7.1	Split Brain Studies	103
4.7.2	The Choice Blindness Paradigm	107
5	Interrogative Theories of Introspection	110
5.1	Introduction	110
5.2	Interrogative Theories of Introspection	114
5.3	Practical Questions and Theoretical Questions	116
5.4	Resolving Practical and Theoretical Questions	120
5.4.1	Resolving Practical Questions About Attitudes	122
5.5	Evaluating the Theoretical Question Theory	127
5.6	Evaluating the Normative Question Theory	132
5.7	Evaluating the Practical Question Theory	137
5.8	Two Problems for the Practical Question Theory	139
5.8.1	Fragile Resolutions	140
5.8.2	Robust Resolutions	142
5.9	Conclusion	144
6	Conclusion	146

CHAPTER 1

INTRODUCTION

This thesis is about self-knowledge, our knowledge of particular facts about our own actions, attitudes, and sensory states. Such knowledge seems to be non-inferential: it does not seem to be arrived at by inference from other things we know or believe. In this respect, it is like much of our knowledge of our immediate environment. But unlike much of our knowledge of our immediate environment, which is a kind of perceptual knowledge, our knowledge of particular facts about our own actions, attitudes, and sensory states seems to be non-perceptual: it does not seem to be based on sensory perception. The problem of self-knowledge is the problem of explaining how we know particular facts about our own actions, attitudes, and sensory states, given that such knowledge seems to be both non-inferential and non-perceptual. In this thesis I develop a solution to the problem of self-knowledge.

I focus, in particular, on our knowledge of particular explanatory facts about our own actions and attitudes. Such knowledge seems to be both non-inferential and non-perceptual, like much of our knowledge of other particular facts about our own actions, attitudes, and sensory states. An important subproblem of the problem of self-knowledge is the problem of explaining how we know particular *explanatory* facts about our own actions and attitudes, given that such knowledge seems to be both non-inferential and non-perceptual. This question remains relatively unexplored in the philosophical literature, since many theorists mistakenly think

that such knowledge is inferential. If such knowledge is inferential, then there is no need to explain how we can know particular explanatory facts about our actions and attitudes in a non-inferential way. I argue this is mistaken and that our knowledge of particular explanatory facts about our own actions and attitudes is non-inferential. This conclusion bears importantly on the question of how we know particular facts about our own actions and attitudes, including, in particular, explanatory facts, given that such knowledge seems to be both non-inferential and non-perceptual. A fully unified answer to this question must explain how we know particular explanatory facts about our actions and attitudes in a non-inferential way.

This thesis consists of four related essays. Essay one and essay four are concerned with the problem of self-knowledge in general. Essay two and essay three are concerned with the problem of explaining how we know particular explanatory facts about our own actions and attitudes. Essay one examines and rejects a general constraint on a solution to the problem of self-knowledge. According to this constraint, a belief is rational only if it is based on sufficient reasons for so believing. Essay two concerns the nature of rational explanations and explanatory language and sets the scene for essay three. Essay three argues that our knowledge of particular explanatory facts about our own attitudes and actions is non-inferential. It develops a theory on which we can come to know particular explanations of our own actions and attitudes by resolving particular practical questions about our own actions and attitudes. Essay four develops this theory as a solution to the problem of explaining how we know particular facts about our own actions and attitudes in general and contrasts it with two related interrogative theories of introspection.

For the remainder of this introduction I will explicate the problem of self-knowledge in more detail along with several salient sub-problems of self-knowledge and connect the four essays which constitute this thesis to the problem and sub-problems. The aim is merely to give an overview of the themes of the essays to follow and to show how they hang together.

1.1 The Problem of Self-Knowledge

The problem of self-knowledge is the problem of explaining how we know particular facts about our own actions, attitudes, and sensory states, given that such knowledge seems to be

both non-inferential and non-perceptual. The presupposition of the problem of self-knowledge is widely, although not universally, accepted. That is, it is widely accepted that our knowledge of particular facts about our own actions, attitudes, and sensory states is both non-inferential and non-perceptual. What this means is that our beliefs about our own actions, attitudes, and sensory states, are not inferred from anything else we believe or know, and that they are not based on states of sensory perception in the way that many of our beliefs about our immediate environment are based on states of sensory perception.

The problem can helpfully be divided into three sub-problems, the problem of explaining how we know particular facts about our own actions, the problem of explaining how we know particular facts about our own attitudes, and the problem of explaining how we know particular facts about our own sensory mental states. And, of course, the problem of explaining how we know our own attitudes can be further divided according to different attitudes. A fully general solution to the problem of self-knowledge would provide a unified solution to all three sub-problems of self-knowledge. Less general solutions would provide a unified solution to any pair of sub-problems. Theorists rarely offer unified solutions to the problem of explaining how we know particular facts about our own actions and the problem of explaining how we know particular facts about our own attitudes.

The sub-problems of explaining how we know particular facts about our own actions and attitudes can be usefully divided into further sub-problems. For example, the sub-problem of explaining how we know particular facts about our own attitudes can be usefully divided into the sub-problems of explaining how we know *positive* facts about our own attitudes, like the fact that I believe that Jane will have another beer, the problem of explaining how we know *negative* facts about our own attitudes, like the fact that I do not believe that Jane will have another beer, and the problem of *explaining* how we know explanatory facts about our own attitudes, like the fact that I believe that Jane will have another beer because she has only had two beers so far.

Most theorists focus on the sub-problem of explaining how we know positive facts about our own attitudes. The solutions they offer to this problem often do not apply to the other sub-problems. Indeed, the solutions they offer to this problem for a particular attitude, often

do not even apply to other attitudes.

In this thesis I develop a unified solution to the problem of explaining how we know particular facts about our own actions and attitudes, including positive, negative, and explanatory facts. I do not pursue the question of whether the solution can be extended to provide a fully unified solution to the problem of self-knowledge. It will be enough to show that the solution is a unified solution to the problem of explaining how we know particular facts about our own actions and attitudes.

A crucial part of my case for the solution I offer will be the sub-problem of explaining how we know particular explanatory facts about our own actions and attitudes. Once we recognise that we must explain how we know particular explanatory facts about our own actions and attitudes if we want a fully unified solution to the problem of self-knowledge a particular theory of self-knowledge emerges as a clear leader.

1.1.1 Reasons and Rationality

What makes the problem of self-knowledge so pressing is that it is hard to see how our beliefs about our own actions, attitudes, and sensory states, can be rational if they are not inferred from other things we know or based on states of sensory perception. One way of capturing this is in terms of the claim that a belief is rational only if it is based on sufficient reasons for so believing. If a belief is not inferred from other things we know or believe or based on some state of sensory perception, then, how can it be based on sufficient reasons for so believing?

It would be a significant constraint on any solution to the problem of self-knowledge if it were true that a belief is rational only if it is based on sufficient reasons for so believing. For then any solution would need to show that our beliefs about our own actions, attitudes, and sensory states are based on sufficient reasons for so believing. Any solutions which didn't would be inadequate. Moreover, if this claim is true, then the problem of self-knowledge can seem pressing, because many of our beliefs about our actions, attitudes, and sensory states, simply do not seem to be based on sufficient reasons for so believing. So a central question for us is whether it is true that a belief is rational only if it is based on sufficient reasons for so believing.

Unfortunately, the question is not particularly clear, since it is not clear what it is for a belief to be based on a reason. If this just means that a belief is inferred from a reason, then it is clearly false that a belief is rational only if it is based on sufficient reasons for so believing. The view would entail that non-inferential knowledge is impossible. But many theorists hold that a belief can be rational without being inferred from sufficient reasons for so believing. On this view, many of our beliefs about our immediate environment arrived at by sensory perception are based on reasons without being inferred from them. I propose to grant that somehow or other the rationality of perceptual beliefs can be explained within this constraint. The question is whether beliefs about our own actions, attitudes, and sensory states, can be.

I call the view which holds that a belief is rational only if it is based on sufficient reasons for so believing *inferentialism*. In essay one I distinguish between two salient kinds of inferentialism to which I offer very different kinds of objections. The first kind holds that a belief is based on a particular fact which is a reason for so believing only if that belief is based on a mental state whose content is a proposition which corresponds to that fact. I call this *moderate inferentialism*. The second kind holds that a belief can be based on a particular fact which is a reason for so believing even though it is not based on a mental state whose content is a proposition which corresponds to that fact. I call this *radical inferentialism*. I argue on the basis of careful consideration of cases that moderate inferentialism is false. I argue that radical inferentialism must give an ad hoc or circular account of what it is for a belief to be based on a reason in order to make the claim that a belief is rational only if it is based on sufficient reasons for so believing come out true. But then the theory is of little interest. So I reject the view that a belief is rational only if it is based on sufficient reasons for so believing, in the sense in which this claim is usually intended.

1.1.2 Knowledge of Rational Explanations

An important, but often overlooked, subproblem of the problem of self-knowledge is that of explaining how we know particular explanatory facts about our own actions and attitudes. We seem to be able to know such explanatory facts in a non-inferential and non-perceptual way. I seem to be able to know, for instance, why I am going to the pub, in a non-inferential

way. I seem to be able to know that I am going to the pub in order to have another beer, in a non-inferential way.

There are, of course, many explanatory facts about our own actions and attitudes which we cannot know in an inferential way. I cannot know in a non-inferential way that I am raising my arm because I am being manipulated by an evil genius. So it is important that we are clear about which kind of explanatory facts can be known in a non-inferential way and which kind of explanatory facts cannot be. On the face of it, we seem to be able to know both reason explanations of our actions and attitudes, explanations expressed by sentences like 'I am going to the pub because Mary is there' and 'I am going to the pub in order to have another beer', and psychological explanations of our actions and attitudes, explanations expressed by sentences like 'I am going to the pub because I believe that Mary is there' and 'I am going to the pub because I want to have another beer'. Both reason explanations and psychological explanations are a kind of rational explanation. Both explanations contrast with merely causal explanations, like the explanations expressed by the sentences 'John is going to the pub because he has a drinking problem' and 'John dropped the plate because he thought that Mary was watching him' (suitably understood).

Essay two is about the relation between reason explanations and psychological explanations, and the relation between rational explanations and merely causal explanations. It is also about the language used to express such explanations. I argue that such language is systematically ambiguous, and that once we realise this we can see that reason explanations are simply a kind of psychological explanation. An important lesson of this essay is that one cannot simply read off the surface form of an explanatory sentence which kind of explanation it expresses. This lesson will be particularly important when it comes to examining claims about knowledge of particular explanations of our own actions and attitudes.

Once we are clear on the distinction between rational explanations and non-rational explanations, or merely causal explanations, we can ask whether our knowledge of particular explanatory facts—i.e. rational explanations—is in fact non-inferential and non-perceptual. And we can ask, if it is, how we know particular explanatory facts about our actions and attitudes, given that such knowledge is non-inferential and non-perceptual. Many theorists

believe that such knowledge is, contrary to appearances, inferential. They are impressed by particular self and other symmetries when it comes to knowledge of particular explanations. The symmetries, if they exist, are impressive. One particular symmetry concerns the similarity between the explanations we give of our own actions and attitudes and the explanations given by suitably placed observers who must make inferences about the explanations of our actions and attitudes based on the evidence they have. We are prone to make the same mistakes as the observers. It is inferred from this that we must be arriving at explanations of our own actions and attitudes by inference too, just like the observers. If so, then our knowledge of particular explanations of our own actions and attitudes is inferential. Many theorists happily accept this argument, and are happy to narrow the scope of the problem of self-knowledge to that of explaining how we know particular non-explanatory facts about our own actions and attitudes. Some theorists have questioned whether the symmetries actually exist. My strategy in this thesis will be to question the inference itself. I argue that the plausibility of the inference from the self and other symmetries depends on an impoverished understanding of the alternative explanations of how we know particular explanatory facts about our own actions and attitudes. It is often tacitly assumed that the only alternative explanation to inference is a kind of inner-sense view of introspection, on which one would not expect such symmetries.

I argue that there is an alternative theory on which our knowledge is non-inferential and which would explain the observed symmetries. According to this theory we arrive at knowledge of particular explanations of our own actions and attitudes by resolving particular practical questions about our own actions and attitudes. For example, I might arrive at knowledge of the fact that I am going to the pub in order to have another beer by resolving the question of whether to go to the pub in order to have another beer. I argue that such a theory predicts the observed symmetries, since this method of arriving at beliefs about the explanations of one's own actions will be prone to the same kind of errors as the method of arriving at beliefs about the explanations of one's own actions by inference against the background assumption that one is rational to some degree.

1.1.3 Interrogative Theories of Introspection

According to the theory developed in essay three, I can arrive at knowledge of particular explanations of my actions and attitudes by resolving particular practical questions about my own actions and attitudes for particular reasons. I can resolve theoretical questions about the explanation of my actions and attitudes by resolving particular practical questions about my own actions and attitudes for particular reasons. This theory is a kind of interrogative theory of introspection, since it holds that we can resolve theoretical questions about our actions and attitudes by resolving related but distinct questions like practical questions about our own actions and attitudes.

Recently, interrogative theories of introspection have provided an alternative to inner-sense theories of introspection. However, there are various kinds of interrogative theory and they are not always clearly distinguished. In essay four, I distinguish between three kinds of interrogative theory of introspection: the practical question theory, the theoretical question theory, and the normative question theory. The theory developed in essay three is a kind of practical question theory. I explicate the key notions of a practical and theoretical question and that of resolving a practical or theoretical question. The framework developed offers a way of thinking about interrogative theories in general. I then evaluate each theory in turn, arguing that the practical question theory is the best interrogative theory of introspection, and that it provides a plausible solution to the problem of explaining how we know particular facts, including negative and explanatory facts about our own actions and attitudes. The question of whether it can provide a fully unified solution to the problem of self-knowledge is left open. I know of no other solution which offers a fully unified solution to the problem of self-knowledge, and the practical question theory provides a unified solution to more subproblems of the problem of self-knowledge than any other theory I know of. This is a good reason for taking the practical question theory seriously as a theory of self-knowledge.

CHAPTER 2

INTROSPECTION WITHOUT INFERENTIALISM

2.1 Introduction

William Clifford famously held that a belief in a particular proposition is rational only if it is based on sufficient reasons for so believing (Clifford, 1877). According to Clifford, a shipowner's belief, say, that his ship is seaworthy is rational only if it is based on sufficient reasons for believing that it is seaworthy. The shipowner's belief is based on sufficient reasons for so believing only if it is based on various facts which, taken together, are sufficient reasons for believing that it is seaworthy. It is based on these facts only if it is inferred from them. Call the view that a belief in a particular proposition is rational only if it is based on sufficient reasons *inferentialism about rationality* or simply *inferentialism*.

Inferentialism is a simple and attractive view about rationality. It is a plausible claim about the kind of beliefs Clifford was interested in, that is, beliefs arrived at by inference alone. But there seem to be many beliefs which can be rational even though they are not based on sufficient reasons for so believing. Beliefs about our immediate environment arrived at by the senses provide one example. At least some of these beliefs seem to be rational even though

they are not based on sufficient reasons for so believing. Beliefs about the past based on memory and basic a priori beliefs arrived at by reflection alone provide two further examples which are a challenge for inferentialism.

Perhaps the most challenging kind of example for inferentialism, however, involves beliefs about our own mental states. Many of these beliefs seem to be rational even though they are not based on sufficient reasons for so believing. My belief that I can feel a pain in my left toe seems to be rational but does not seem to be based on sufficient reasons for so believing. My belief that I believe that Wagner was a better composer than Verdi seems to be rational but does not seem to be based on sufficient reasons for so believing.

There are two ways of qualifying inferentialism in response to these problems. One way, which I think is Clifford's own view, is to restrict it to beliefs arrived at by inference alone. It is very plausible that a belief arrived at by inference is rational only if it is based on sufficient reasons for so believing. If one qualifies inferentialism in this way then one needs another story about the rationality of beliefs which are not arrived at by inference.

Another way of qualifying inferentialism, which many find attractive, is to reject the view of what it is for a belief to be based on sufficient reasons which makes for a conflict between inferentialism and the examples. This way of responding holds that there is something deeply right about the claim that a belief is rational only if it is based on sufficient reasons. Indeed, the claim can sound like a truism. What generates the problem, according to this response, is the view of what it is for a belief to be based on sufficient reasons for so believing. Perhaps a belief can be based on a fact which is a reason for so believing even though it is not *inferred* from that fact. Perhaps what we need is *inferentialism without inference*.

There have been various attempts by those sympathetic to inferentialism to explain how beliefs about our immediate environment based on the senses can be based on sufficient reasons without being inferred from them (Brewer, 1999; Schroeder, 2011b, 2015). Perhaps those beliefs are based on the very facts about the immediate environment they are about, and so based on excellent reasons for believing. Perhaps those beliefs are based on facts about one's perceptual experiences themselves, without being inferred from them, and therefore plausibly based on sufficient reasons for so believing. This corresponds to a traditional

way of accommodating perception within inferentialism, which appeals to facts about how things appear to one. Such views of the epistemology of perception demonstrate the grip of inferentialism on our epistemological thinking.

The aim of this essay is to argue that inferentialism cannot account for the rationality of beliefs about our own mental states. While inferentialism can be made to look at least *prima facie* plausible in the case of perception, it cannot be made to look plausible in the case of introspection, broadly construed. Whereas those sympathetic to inferentialism as a general theory of the rationality of belief are keen to argue for a kind of *inferentialism without inference*, I will be arguing for a kind of *introspection without inferentialism*.

Here is how I will proceed. In the next section, §2, I will further clarify and motivate inferentialism. It will be important to be clear about what inferentialism is, for once inferentialism is severed from inference it is a notoriously difficult thesis to evaluate. I distinguish between two kinds of inferentialism which will be of central interest to us and to which I offer very different kinds of objections. The first kind holds that a belief is based on a particular fact which is a reason for so believing only if the following condition is met: the belief is based on a mental state whose content is a proposition which corresponds to the fact. (The difference between the expression ‘based on’ in the antecedent and the consequent will be discussed below). I call this moderate inferentialism. The second kind holds that a belief can be based on a particular fact which is a reason for so believing even though it is not based on a mental state whose content is a proposition which corresponds to that fact. I call this radical inferentialism. Now, moderate inferentialism stakes out a claim which can obviously be evaluated. Moreover, it can be made to sound like a plausible way of developing inferentialism without inference. In §3, I argue on the basis of careful consideration of cases that our beliefs about our own attitudes can be rational even though they are not based on sufficient reasons for so believing in the sense that moderate inferentialism holds they must. Radical inferentialism allows that a belief can be based on a fact without being psychologically based on a corresponding psychological state and thus avoids the objections raised against moderate inferentialism. However, it faces objections of its own. In §4, I argue that any semblance radical inferentialism has to being a theory of the rationality of belief is based on various

confusions about what it is for a belief to be based on a fact, or for a belief to be based on another mental state. Proponents of radical inferentialism attempt to squeeze introspection into an epistemological shoe that it simply will not fit.

2.2 Inferentialism

Inferentialism is a simple and attractive picture of the rationality of belief. How could anyone deny that a belief is rational only if it is based on sufficient reasons for so believing? Indeed, if we focus on beliefs arrived at by inference, inferentialism has a lot going for it. Let's begin, then, by getting clearer on *restricted inferentialism*, understood initially as a view about the rationality of beliefs arrived at by inference alone. We will then ask how it must be modified in order to account for the rationality of all beliefs, especially beliefs about our own mental states.

Inferentialism concerns the relation between rational belief and reasons for believing. The relevant sense of rational belief here is one intimately connected with knowledge. A particular belief amounts to knowledge only if it is rational. More carefully: you know some fact only if you rationally believe a proposition which corresponds to the fact. (In speaking of a proposition corresponding to a fact, I mean to avoid the controversy concerning whether true propositions and facts stand in a one to one relation or a many to one relation). Perhaps there is more than one sense in which a belief can be rational. Then inferentialism concerns the sense of rationality connected with knowledge in this way, or at least concerns the sense of rationality which has some prospect of being connected with knowledge in this way. Your belief that the gods love you does not meet the conditions for knowledge if it is rational only in the sense that unless you believe that the gods love you you will suffer eternal punishment, and it is rational for you to do anything you can to avoid eternal punishment. Inferentialism holds that a belief is rational only if it is based on sufficient reasons. That is, a belief is rational only if it is based on various facts which, taken together, are sufficient reasons for so believing. What are reasons for believing? And what is it for various facts, taken together, to be sufficient reasons for believing something? Reasons for believing are facts which support or favour believing. Not much more can be uncontroversially said about them. The fact that

Wagner had a greater harmonic range than Verdi is a reason for believing that he was a better composer than Verdi. The fact that the icecaps are melting is a reason for believing that global temperatures are rising.

Let's call reasons for believing in this sense *normative* reasons. We can contrast reasons for believing in the normative sense with reasons for believing in the *rationalizing* sense. When some mental state makes it rational for you to believe some proposition, we might say that it is a reason for believing that proposition in the rationalizing sense. My belief that Wagner had a greater harmonic range than Verdi is a reason for me to believe that Wagner was a better composer than Verdi in the rationalizing sense of 'reason for believing. Notice that the fact that Wagner had a greater harmonic range than Verdi itself does not itself make it rational in some sense and to some degree for me to believe that Wagner was a better composer than Verdi. Of course, my being aware of this fact, may make it rational to so believe. As we will see later, radical inferentialism can look plausible if these two senses of 'reason for believing' are confused. It is important to keep them apart. In order to do so, whenever I speak of reasons for believing in an unqualified way, I usually mean reasons in the sense that the fact that Wagner had a greater harmonic range than Verdi is a reason for believing that he was a better composer than Verdi. I will speak of reasons for believing in the rationalizing sense when I must, where by 'rationalizing', I just mean, explaining why it is rational in some sense and to some degree.

What is it for various facts, taken together, to be sufficient reasons for believing something? We have an intuitive understanding of sufficiency. And we won't need much more than this for the discussion to follow. Many theorists believe that sufficiency of reasons is a matter of the weight of reasons *for* and the the weight of reasons *against* (Schroeder, 2011a, 2012). One might have sufficient reasons for believing some proposition if and only if the reasons one has for believing that proposition are at least as good as the reasons one has against believing it. One might have sufficient reasons for believing some proposition if and only if the reasons one has for believing that proposition are significantly better than the reasons one has against believing it. We needn't take a stand. But it is important to keep in mind that it does not follow that if one adds reasons to a set of reasons which are sufficient for believing some

proposition, that set of reasons will remain sufficient. Adding a reason to a set of reasons might change the weights of the existing reasons, or might change the overall balance by weighing against.

This much is part of a widely agreed upon picture of rationality and reasons. Our evaluation of inferentialism will not depend on questioning any of these assumptions. Rather, it is going to depend crucially on what it is for a belief to be *based on* sufficient reasons for believing, if this isn't for the belief to be inferred from sufficient reasons for so believing.

2.2.1 Factual and Psychological Basing

For a belief to be *based on* a particular reason for believing is for it to be based on a fact which is a reason for so believing. My belief that Wagner was a better composer than Verdi is based on the fact that Wagner had a greater harmonic range. The fact that Wagner had a greater harmonic range is a reason for believing that Wagner was a better composer than Verdi. So my belief that Wagner was a better composer than Verdi is based on a reason. My belief is based on a reason in this case because I inferred that Wagner was a better composer than Verdi from the fact that he had a greater harmonic range than Verdi. Inferring one fact from another fact is a way, although not necessarily the only way, for one's belief in some proposition to be based on a fact.

Just as a belief can be based on a fact, it can be based on another belief. My belief that Wagner was a better composer than Verdi is based on my belief that Wagner had a greater harmonic range than Verdi. The preposition 'based on', then, seems to be able to be used to express a relation between a mental state and a fact, as when one asserts that my belief is based on the fact that Wagner had a greater harmonic range than Verdi, and between a mental state and another mental state, as when one asserts that my belief is based on my belief that Wagner had a greater harmonic range than Verdi. Arriving at one belief by inference from another belief is one way, although not necessarily the only way, for one belief to be psychologically based on another. Since these relations are distinct, it will be important to distinguish them. Although it is always possible to tell which relation is involved in terms of the relata involved, in order to avoid potential confusion I will use 'based on' to express

a relation between a mental state and a fact, and ‘psychologically based on’ to express a relation between a mental state and another mental state.

Although the relation of being based on a fact and of being based on a mental state are distinct relations, they seem to be related in an important way. My belief that Wagner was a better composer than Verdi is based on the fact that Wagner had a greater harmonic range than Verdi, it seems, only if my belief that Wagner was a better composer than Verdi is psychologically based on my belief that Wagner had a greater harmonic range than Verdi. However, the converse does not hold: if my belief is psychologically based on my belief that Wagner had a greater harmonic range than Verdi, it does not follow that my belief is based on the fact that Wagner had a greater harmonic range than Verdi. Wagner might not have had a greater harmonic range than Verdi. Or he might have, but I might not have known that he did.

Our ordinary notion of the relation of being based on a fact seems to hold that what is true of my belief about Wagner is true in general: a belief is based on a fact only if it is psychologically based on a *belief* in a corresponding proposition. But it is this claim that the sophisticated inferentialist is keen to reject. They want to allow that a belief can be based on a fact without being psychologically based on a belief in a corresponding proposition.

2.2.2 *Conservative, Moderate, and Radical Inferentialism*

We can distinguish between conservative, moderate, and radical inferentialism in terms of how they differ with respect to what they hold that it takes for a belief to be based on a reason. According to conservative inferentialism, a belief is based on a fact only if it is psychologically based on a *belief* in a corresponding proposition. Moderate inferentialism rejects conservative inferentialism’s claim that a belief is based on a fact only if it is psychologically based on a *belief* in a corresponding proposition. It holds that the truth in conservative inferentialism is that *a belief is based on a fact only if it is psychologically based on a mental state whose content is a proposition which corresponds to the fact*. Radical inferentialism holds that it is possible for a belief to be based on a fact without being psychologically based on a mental state whose content is a proposition which corresponds to the fact. It might do this by

rejecting the requirement that for a belief to be based on a fact it must be based on a mental state at all, or it might hold that while it must be based on a mental state, it need not be based on a mental state whose content is a proposition which corresponds to the fact.

It is the distinction between conservative, moderate, and radical inferentialism, which will be of central importance in what follows. But we should note at least one further dimension of variation among different kinds of inferentialism which matters for the prospects of inferentialism in general. We can distinguish between subjective and objective versions of inferentialism according to whether it is subjective reasons or objective reasons which are taken to be relevant. According to subjective inferentialism a belief is rational only if it is based on sufficient subjective reasons for so believing (Schroeder, 2008). According to objective inferentialism a belief is rational only if it is based on sufficient objective reasons for so believing. A subjective reason is a proposition which has the following property: it is an objective reason if it is true. The proposition that Wagner wrote better counterpoint than Verdi is a subjective reason for believing that Wagner was a better composer than Verdi. Subjective inferentialism must replace the conception of a belief being based on a fact with the corresponding notion of a belief being based on a proposition. It might hold that a belief is based on a proposition only if it is psychologically based on a belief in that proposition, or at least on a mental state whose content is that proposition. Since the difference between subjective and objective inferentialism will not matter for our purposes in what follows, I will continue to write in terms of objective inferentialism and in terms of beliefs being based on facts. It will be obvious that what I say against objective inferentialism of various stripes also applies against subjective inferentialism of various stripes. One final proviso. Sometimes theorists use the expression ‘subjective reason for believing’ to refer to what I have called reasons for believing in the rationalizing sense above. This is *not* what I mean by subjective reasons for believing. Subjective reasons for believing, like objective reasons for believing, do not explain why it is rational in some sense and to some degree for someone to believe something. It is the *having* of subjective or objective reasons which does so. The distinction between having and not having a reason matters just as much for subjective reasons as we are understanding them as it does for objective reasons.

2.3 The Case Against Moderate Inferentialism

Moderate inferentialism is an ambitious thesis. It holds that a belief is rational only if it is based on sufficient reasons for so believing and that a belief is based on a particular fact/reason only if it is psychologically based on a mental state whose content corresponds to that fact/reason. The best argument against inferentialism simply involves offering cases which could plausibly obtain where a belief about our own mental states is rational but it does not seem to be based on sufficient reasons for so believing (Boghossian, 1989, pp. 7-8; Gallois, 1996, p. 38; Goldman, 2006, p. 230). Upon being presented with such a case, the proponent of moderate inferentialism can argue either that the case given is one in which the belief is not rational or the case is one in which the belief *is* based on sufficient reasons, or the case is one in which it is neither rational nor based on sufficient reasons.

Beliefs about our own actions, our own attitudes, and our own sensory states, provide a challenge for moderate inferentialism, for many such beliefs seem to be rational but do not seem to be based on sufficient reasons for so believing. My belief that I will have another beer—a belief about a future action of mine—does not seem to be based on sufficient reasons for so believing. But it seems to amount to knowledge, and so must be rational. My belief that I can feel a moderate pain in my back—a belief about a sensory state of mine—does not seem to be based on sufficient reasons for so believing. But it too seems to amount to knowledge, and so must be rational. Finally, my belief that I believe that Hamburg is west of Berlin—a belief about an attitude of mine—does not seem to be based on sufficient reasons for so believing. But it seems to amount to knowledge, and so must be rational. The moderate inferentialist is committed to denying each of these claims.

I am going to set aside beliefs about actions and beliefs about sensory states in what follows. While these provide interesting challenges to moderate inferentialism, I am going to focus on the case of beliefs about our own attitudes. The reason for this is that beliefs about our own attitudes seem to provide a case where both inferential and perceptual models of how we arrive at such beliefs seem the most implausible. In contrast, many theorists find inferential models of how we arrive at our beliefs about our own actions plausible (Paul, 2009a), and many theorists find perceptual models of how we arrive at beliefs about our own

sensory states plausible (Carruthers, 2011). Inferential and perceptual models are amenable to moderate inferentialism. By focussing on the case of beliefs about our own attitudes, we need not discuss controversial issues surrounding perceptual models, since it is widely assumed that perceptual models of how we arrive at our beliefs about our own attitudes are implausible.

Discussions of beliefs about our own attitudes tend to focus on a few central attitudes like belief, desire, and intention. In what follows, I will follow this tradition. However, I am going to focus not only on positive beliefs about our own attitudes, but also negative beliefs (Sosa, 2003; Stoljar, 2012) and explanatory beliefs (Pryor, 2005; Setiya, 2013; Baker, 2015). My belief that I believe that Hamburg is west of Berlin—a positive belief about an attitude of mine—does not seem to be based on sufficient reasons for so believing. Moreover, my belief that I do *not* believe that Hamburg is south of St. Petersburg—a negative belief about an attitude of mine—does not seem to be based on sufficient reasons for so believing. And, finally, my belief that I believe that Hamburg is west of Berlin because I believe that Potsdam is west of Berlin and that Hamburg is west of Potsdam—an explanatory belief about an attitude of mine—does not seem to be based on sufficient reasons for so believing. Although I do not think that the focus on a few central attitudes like belief, desire, and intention has had a distorting effect on discussions of self-knowledge, I do think that the narrow focus on *positive* beliefs about our attitudes has had such an effect.

We can begin by examining an example of a positive belief about a belief. I believe that Wagner was a better composer than Verdi. I believe that I believe that Wagner was a better composer than Verdi. I didn't believe this a moment ago, since I hadn't considered the question of whether I believe that Wagner was a better composer than Verdi before. Do I have sufficient reasons for believing that I believe that Wagner was a better composer than Verdi? On the face of it, I do not. I may have reasons for believing that Wagner was a better composer than Verdi. The fact that Wagner had a greater harmonic range than Verdi is a reason I have for believing that he was a better composer than Verdi, for example. But such reasons are not typically reasons for believing that I *believe*. The fact that Wagner had a greater harmonic range than Verdi isn't a reason for me to believe that I believe that Wagner was a better composer than Verdi.

In asking whether I have sufficient reasons for believing that I believe that Wagner was a better composer than Verdi, it is worth asking what kind of facts could be reasons for believing that I believe this. Perhaps I recently asserted that Wagner was a better composer than Verdi. The fact that I recently asserted that Wagner was a better composer than Verdi is a reason I have—assuming that I believe that I recently asserted that Wagner was a better composer than Verdi—for believing that I believe that Wagner was a better composer than Verdi. But I can assure you that I didn't recently assert this. So this can't be a reason I have for believing it. Suppose I was recently given a choice between listening to the music of Wagner or the music of Verdi, and I chose to listen to the music of Wagner. This could be a reason I have for believing that I believe that Wagner was a better composer than Verdi. But, again, I can assure you that I wasn't recently given such a choice. Or suppose I just heard a musicologist assert that Wagner was a better composer than Verdi. This could be a reason I have for believing that Wagner was a better composer than Verdi. Again, I can assure you that I didn't just hear a musicologist assert that Wagner was a better composer than Verdi. So, it seems, although I know that I believe that Wagner was a better composer than Verdi, and so rationally believe it, my belief is not based on sufficient reasons for so believing.

As we have seen there are various reasons I could have for believing that I believe that Wagner was a better composer than Verdi, but when we consider the case at hand, we see that I do not seem to have any of these reasons, and so do not seem to have sufficient reasons at all for believing that I believe that Wagner was a better composer than Verdi. Many theorists are convinced by reflection on cases like this that our beliefs about our own beliefs can be rational without being based on sufficient reasons for so believing. But others are not convinced. In considering the example above, we focused on what we might call 'public reasons' for believing that I believe that Wagner is a better composer than Verdi. Two of the reasons were facts about the effects of a belief like the belief that Wagner is a better composer than Verdi: they were facts about assertion and facts about choice. One of the reasons was a fact about a cause of beliefs like the belief that Wagner is a better composer than Verdi. These are the kind of facts a third party might appeal to in attributing the belief that Wagner was a better composer than Verdi to me. But in considering only public reasons, we risk ignoring what we

might call ‘private reasons’. Some theorists hold that in cases like the one just considered, we have private reasons for so believing, and so may well have sufficient reasons for so believing, when we take all reasons into account.

Private reasons include facts like facts about how I currently feel, facts about what I am currently saying to myself in inner-speech, or have just said to myself in inner-speech, and facts about what I am currently imagining (Ryle, 1949; Lormand, 1996; Lawlor, 2009; Carruthers, 2011; Cassam, 2014). Suppose when I consider the question of whether I believe that Wagner was a better composer than Verdi, I hear myself say in inner-speech that Wagner was a better composer than Verdi. The fact that I have just said to myself in inner-speech that Wagner was a better composer than Verdi is a reason for believing that I believe that Wagner was a better composer than Verdi. This is a reason I could have for believing that I believe that Wagner was a better composer than Verdi. I can assure you, however, that I didn’t hear myself saying in inner-speech that Wagner was a better composer than Verdi. I heard myself ask myself whether I believe that Wagner was a better composer than Verdi in inner-speech. The next thing I heard in inner-speech was that I believe that Wagner was a better composer than Verdi. Of course, if I had asked myself in inner-speech whether Wagner was a better composer than Verdi, then I might have heard myself in inner-speech say that Wagner was a better composer than Verdi. Then I would have had a good reason for believing that I believe that Wagner was a better composer than Verdi. But I can assure you that I didn’t. And, if you reflect on your own inner-speech when the question you consider is whether you *believe* some proposition, you will agree.

Focusing on the case of positive beliefs about beliefs threatens to distort our view of the facts. Theorists sympathetic to moderate inferentialism like to point out that the fact that one has just asserted, or said, in inner-speech is a reason for one to believe that one believes what one has just asserted. Suppose we agree. I have just argued that in many cases what we assert or say in inner speech after considering some question about our own beliefs is not that something is the case but rather that we believe that it is the case. But once we move beyond the case of positive beliefs about beliefs, the connection between particular sentences heard in inner-speech and one’s attitudes becomes more strained. Suppose I consider whether I intend

to have another beer. I might hear myself ask this question in inner-speech. What might I hear in response? If anything, it will plausibly be that I intend to have another beer. I might hear myself say in inner-speech that I should have another beer, or that having another beer would be nice. Then I would have reasons for believing that I intend to have another beer, although, admittedly, not particularly strong ones. But it is only in rare cases that this is what I hear.

A favourite case of those theorists sympathetic to moderate inferentialism is that of desire (Lawlor, 2009). It is easy to see why. Very often we have to find out what we want by inference. It is a matter of discovery. Our beliefs about many of our desires are based on sufficient reasons for so believing. But many of our beliefs about our desires are not based on sufficient reasons for so believing. This is so for so-called instrumental desires. Suppose I want to get to Hamburg, and know that in order to get to Hamburg I have to buy a train ticket. I want to buy a train ticket. I know that I want to buy a train ticket. But I do not have sufficient reasons for believing that I want to buy a train ticket. Again, it is not that I couldn't have such reasons. It is that I don't. Consider the case in more detail. I have just decided to go to Hamburg. I consider what I must to do get there. I know that in order to get to Hamburg I have to buy a train ticket. I immediately come to desire to buy a train ticket. None of this is accompanied by any kind of inner speech or feeling. My thought is directed at the world, at Hamburg, at train tickets, etc. Then I consider the question of whether I want to buy a train ticket. (Perhaps a stranger stops me on the street and asks). I have no reasons for believing that I want to buy a train ticket, public or private. But I immediately come to believe that I want to buy a train ticket. There is no reason to think that my belief will not amount to knowledge, and therefore be rational, even though it is not based on sufficient reasons for so believing.

So far I have focused on positive beliefs about attitudes. A good case can be made against moderate inferentialism here. A better case can be made when we consider negative beliefs and explanatory beliefs, since negative beliefs and explanatory beliefs stand in a different relation to inner-speech. Consider negative beliefs first. Suppose I do not believe that Hamburg is at a higher elevation than Berlin. I don't know whether it is. Suppose I consider the question of whether I believe that Hamburg is at a higher elevation than Berlin. I come to believe that I do not believe that Hamburg is at a higher elevation than Berlin. What reasons could I have

for believing that I do not believe that Hamburg is at a higher elevation than Berlin? Well, the fact that I have never asserted that it is or isn't is not a reason I have for believing that I do not believe it. There are ever so many things I haven't said out loud which I do believe. It is sometimes suggested that I might hear myself in inner-speech say "no" in response to the question of whether Hamburg is at a higher elevation than Berlin, and so come to know that I do not believe that Hamburg is at a higher elevation than Berlin. But that won't help in this case. I won't say "no": I don't know whether it is. Nor will it help if I say "I don't know" in inner-speech, for that presupposes that I know that I do not know.

Finally, consider the case of explanatory beliefs about one's own attitudes. Suppose I believe that Hamburg is west of Berlin in part because (I believe) Hamburg is west of Potsdam, and suppose I consider the question of whether I believe that Hamburg is west of Berlin because Hamburg is west of Potsdam. Immediately I come to believe that I believe that Hamburg is west of Berlin because Hamburg is west of Potsdam. There are various reasons I could have for believing this. Indeed, perhaps the fact that I believe that Hamburg is west of Berlin, and the fact that I believe that Hamburg is west of Potsdam, and the fact that the fact that Hamburg is west of Potsdam is a reason for believing that Hamburg is west of Berlin, are, together, sufficient reasons for believing it. But I have never considered the question of whether the fact that Hamburg is west of Potsdam is a reason for believing that Hamburg is west of Berlin. That is a rather sophisticated thing to think. So this isn't among my reasons. And without this fact being among my reasons, I do not seem to have sufficient reasons for believing that I believe that Hamburg is west of Berlin because (I believe) Hamburg is west of Potsdam. But there is no reason to think that my explanatory belief will not amount to knowledge, and therefore be rational, even though it is not based on sufficient reasons for so believing.

In this section I have argued, on the basis of careful reflection on particular cases, that our beliefs, positive, negative, and explanatory, about our own attitudes often amount to knowledge, and so are rational, even though they are not based on sufficient reasons for so believing. This is because in many cases, one simply does not have sufficient reasons for so believing, and so cannot believe on the basis of sufficient reasons for so believing. I have examined var-

ious facts which could be reasons we have for so believing and have argued that although we sometimes have such reasons, we often do not. Focusing on the case of positive beliefs about our own beliefs or on the case of positive beliefs about our own desires suggest that we often have such reasons for believing. But it is a mistake to focus on these cases, as it obscures our view of the facts.

In pursuing the argument above, we allowed that it would be rather unproblematic for one to *have* the particular reasons in question, if they were to obtain. For instance, we allowed that *if* I had said in inner speech that Wagner was a better composer than Verdi I would *have* the fact that I had said in inner speech that Wagner was a better composer than Verdi among my reasons for believing. But remember that according to moderate inferentialism, a belief is based on a fact only if it is psychologically based on a mental state whose content is a proposition which corresponds to that fact. So, according to moderate inferentialism, for my belief that I believe that Wagner was a better composer than Verdi to be based on the fact that I have said in inner speech that Wagner was a better composer than Verdi, my belief would have to be psychologically based on a mental state whose content is the proposition that I have said in inner speech that Wagner was a better composer than Verdi. What this brings out is that it is one thing to say something to oneself in inner speech that Wagner was a better composer than Verdi, and quite another thing to have the fact that one has said to oneself in inner speech that Wagner was a better composer than Verdi among one's reasons for believing, and for it to be something one's belief might be based on. Theorists sympathetic to inferentialism often speak as if there were no gap between a reason obtaining and it being available as something one's belief might be based on. But it is a further question as to whether the reason is available in this sense. As we will see below, according to radical inferentialism, there is no such gap when it comes to private reasons. But it is important at this point to see that this is something which distinguishes between moderate and radical inferentialism, and that moderate inferentialists are not entitled to assume that there is no such gap.

2.4 The Case Against Radical Inferentialism

The structure of the case against moderate inferentialism was straightforward. This is because moderate inferentialism is committed to a claim which can be tested against cases. The case against radical inferentialism will be different, since radical inferentialism, as we have defined it, merely denies one part of moderate inferentialism. Radical inferentialism attempts to hold on to the view that a belief is rational only if it is based on sufficient reasons for so believing while rejecting the view that a belief is based on a reason only if it is psychologically based on a mental state whose content is a proposition which corresponds to that reason.

2.4.1 Cassam's Radical Inferentialism

In his recent book Quassim Cassam claims to be defending an inferential theory of self-knowledge (Cassam, 2014). Cassam doesn't have inferentialism as we are understanding it in mind, but his theory is amenable to inferentialism. Indeed, Cassam advertises his theory in such a way that it sounds like he is defending the view that beliefs about our own mental states are rational only if they are based on sufficient reasons, where to be based on sufficient reasons the belief must be inferred from them. Cassam thus seems to be defending conservative inferentialism. But closer inspection reveals that this is not his view. Cassam seems to allow that a belief can be based on some fact even if it is not psychologically based on a mental state whose content is a proposition which corresponds to that fact. So Cassam seems to be committed to rejecting even moderate inferentialism. Perhaps this isn't Cassam's considered view. But it is certainly a view suggested by his writing. In any case, there is clearly a view like this in the vicinity worth considering. (Sometimes it sounds as though Cassam is defending the view that our beliefs about our mental states are always at least partly psychologically based on other beliefs. Such a view is compatible with the denial of inferentialism as we are understanding it, since such beliefs may nonetheless not be based on sufficient reasons for so believing).

The view I have in mind makes a distinction between what it takes for a belief to be based on a reason when that reason is a 'public reason', and what it takes for a belief to be based on a reason when that reason is a 'private reason'. Cassam suggests something along these

lines when he writes: “If E is your evidence that you have a particular attitude A, and you know that you have A by inference from E, then you need access to E. What kind of access to E? That depends on the nature of your evidence. If your evidence is behavioural then your access to E might have to be perceptual, but this won’t be true if your evidence consists of your passing thoughts or inner speech” (Cassam, 2014, p. 139). So in the case of ‘public reasons’, in order for a belief to be based on a reason it must be psychologically based on a mental state whose content corresponds to that reason. In the case of ‘private reasons’ a belief is based on the fact that you are in some mental state if it is psychologically based on that mental state. Suppose that you are feeling a particular way, and the fact that you are feeling a particular way is a reason for you to believe that you want to listen to a Wagner opera. On this view, there are two ways for your belief that you want to listen to a Wagner opera to be based on this reason. It can be psychologically based on a mental state whose content is a proposition which corresponds to the reason—as it would be if it were based on a belief that you feel the way you do—or it can be psychologically based on your feeling the way you do itself. Either way, on this view, it is based on the fact that you feel the way you do. And if this is a reason for believing that you want to listen to a Wagner opera, then it is based on a reason for believing that you want to listen to a Wagner opera.

Perhaps we can tolerate the disjunctive analysis of what it is for a belief to be based on a reason offered by this view. But that seems to be the least of the problems for this view. The central problem for this kind of radical inferentialism is that beliefs formed in this way would be manifestly irrational. Let me explain. We have already distinguished two senses of ‘reason for believing’, the normative sense and the rationalizing sense. Now, typically, when a belief is based on a reason for believing in the normative sense of ‘reason for believing’, it is also psychologically based on a reason for believing in the rationalizing sense of ‘reason for believing’. So my belief is based on the fact that Wagner had a broader harmonic range than Verdi, and it is psychologically based on my belief that Wagner had a broader harmonic range than Verdi. The former fact is a reason for believing that Wagner was a better composer than Verdi, the latter mental state is a reason for believing that Wagner was a better composer than Verdi in the sense that it is something which explains why it is rational for me to believe

this. Radical inferentialism breaks this link. My feeling a particular way does not explain why it is rational for me to believe that I want to listen to a Wagner opera. If my belief is psychologically based on this feeling, then it isn't based on a reason for believing that I want to listen to a Wagner opera in the rationalizing sense. What would explain why it is rational for me to believe that I want to listen to a Wagner opera is my being in a mental state whose content is a proposition which corresponds to the fact that I am feeling the way I am. And if my belief is psychologically based on this mental state, then it is based on a reason for believing that I want to listen to a Wagner opera in the rationalizing sense and so is plausibly rational. It is important that we do not slide between these two claims. In general, it does not follow that if being in a particular mental state with a content which corresponds to a particular fact about one's mental states explains why it is rational to believe something or other, then the mental state the fact is about explains why it is rational to believe that thing. Any plausibility that this kind of radical inferentialism has results from sliding between the two.

2.4.2 Zimmerman's Radical Inferentialism

We finally come to a discussion that we have been postponing throughout, and that is a discussion of what we might call simple radical inferentialism. According to simple radical inferentialism, our beliefs about our own mental states are based on the facts about our mental states themselves but they are not psychologically based on a mental state whose content is a proposition which corresponds to that fact. My belief that I believe that Wagner was a better composer than Verdi is based on the fact that I believe that Wagner was a better composer than Verdi. Now, the fact that I believe that Wagner was a better composer than Verdi is an excellent reason for me to believe that I believe that Wagner was a better composer than Verdi. If my belief is based on this fact, then it is based on sufficient reasons for so believing. But this is a kind of radical inferentialism. It holds that my belief can be based on the fact that I believe that Wagner was a better composer than Verdi even though it is not psychologically based on a mental state whose content is a proposition which corresponds to the fact that I believe that Wagner was a better composer than Verdi.

Aaron Zimmerman has defended a version of this kind of radical inferentialism (Zimmerman, 2006, 2008; Shoemaker, 2012). Zimmerman is sensitive to the idea that in order for one's belief to be based on a reason, one needs to stand in some kind of relation to that reason. He is aware that in order for my belief that Wagner was a better composer than Verdi to be based on the fact that Wagner had a broader harmonic range than Verdi, for instance, I must stand in some kind of relation to that reason. Indeed, plausibly, I must believe a corresponding proposition. But Zimmerman thinks that we need to weaken the relation for beliefs about our own mental states. The most general relation we stand in to reasons for believing, according to Zimmerman, is the relation of availability. In order for a belief to be based on a reason, that reason must be available. In the case of beliefs about the non-mental world, such reasons are available because we are in mental states whose contents are propositions which correspond to such reasons. Zimmerman suggests that the fact that we are in a particular mental state is available just in case that mental state itself is *available* or *access conscious*—that is, poised to play particular roles in our mental economy. Zimmerman writes: “when a person believes that *p* in this full-blown manner, the fact that she believes that *p* will be *available* to her for use in second-order reasoning” (Zimmerman, 2008, p. 336). If my belief is psychologically based on such an available or access conscious mental state, it is based on the fact that I am in that mental state. As we can see, this is a kind of radical inferentialism. It allows a belief to be based on a fact without being psychologically based on a mental state whose content is a proposition which corresponds to that fact. Why the restriction to mental states which are available or access conscious? Because Zimmerman is worried about the objection that a belief can be based on some fact only if that fact is ‘available’ in some sense or other. The objection is a good one. It is hard to see how a belief can be based on a fact unless that fact is somehow ‘available’ to the agent. Obviously if the agent is in a mental state whose content is a proposition which corresponds to that fact, then the fact is ‘available’ in some sense. But Zimmerman trades here on an ambiguity in ‘available’ and ‘accessible’. When a mental state is available or access conscious, then its content is available or accessible. It does not follow that the fact that you are in that mental state is available or accessible. So Zimmerman hasn't replaced moderate inferentialism with a plausible alternative which holds that a belief

is based on a fact only if is psychologically based on a proposition corresponding to that fact's being available to the agent. His way of developing radical inferentialism fails to meet the intuitive objection. But even supposing that it could meet this objection, this view runs into deeper problems.

We cannot object to Zimmerman's radical inferentialism in the way that we objected to Cassam's, since Zimmerman's version of radical inferentialism does not obviously sever the link between being based on a reason in the normative sense and being based on a reason for believing in the rationalizing sense. If my belief that Wagner was a better composer than Verdi is psychologically based on my belief that Wagner was a better composer than Verdi, then it is arguably based on a reason for believing in the rationalizing sense.

My main objection to Zimmerman's radical inferentialism is this. Not only is there the connection between being based on a reason in the normative sense and being based on a reason in the rationalizing sense which we have already noted, but there is a further connection. When a belief is based on a reason in the normative sense, and so based on a reason in the rationalizing sense, it is usually the case that the mental state which is a reason in the rationalizing sense is a reason *because* its content is a proposition which corresponds to a reason in the normative sense. We answer the question of *why* the mental state rationalizes believing what it does in terms of the fact that its content corresponds to a reason for believing. Let's ask whether this explanation obtains here. Let's ask whether my belief that Wagner is a better composer than Verdi is a reason for believing that I believe that Wagner is a better composer than Verdi because its content corresponds to a reason for believing that I believe that Wagner is a better composer than Verdi. The answer to this question is pretty clearly 'no'. The content of my belief is the proposition that Wagner is a better composer than Verdi. This does not correspond to a fact which is a reason for me to believe that I believe that Wagner is a better composer than Verdi. Nonetheless, it is plausible that my belief that Wagner is a better composer than Verdi is a reason for me to believe that I believe that Wagner is a better composer than Verdi in the rationalizing sense. Its being so must have a different explanation. At this point the question arises: if my belief that I believe that Wagner is a better composer than Verdi is based on my belief that Wagner is a better composer than Verdi itself, and this

is a reason in the rationalizing sense for me to believe that I believe that Wagner is a better composer than Verdi, then won't my belief be rational? The answer to this is plausibly 'yes'. But since we have already seen that whenever a belief is based on a reason in the normative sense it is psychologically based on a reason in the rationalizing sense, shouldn't we conclude that it is being based on a reason in the rationalizing sense which is relevant to the rationality of belief, and not being based on a reason in the normative sense? The thought is that claiming that my belief that I believe that Wagner is a better composer than Verdi is based on the fact that I believe that Wagner is a better composer than Verdi and that it is therefore based on sufficient reasons in the sense relevant to inferentialism, is an idle wheel. We are simply trying to fit beliefs about our own mental states into an epistemological shoe that they will not fit.

In effect, I have just argued that if there is a sense in which a belief is rational only if it is based on sufficient reasons, then this sense is the sense of being psychologically based on reasons in the rationalizing sense. It is this sense which is in play when theorists of self-knowledge rail against 'no reasons' views of self-knowledge (Peacocke, 1999; Zimmerman, 2006; Dretske, 2011; Fernandez, 2013). But inferentialism claims that a belief is rational only if it is based on sufficient reasons in a different sense. Any plausibility that Zimmerman's form of radical inferentialism has is derived from sliding between the two.

2.5 Conclusion

There is widespread agreement among theorists of self-knowledge that our beliefs about our own mental states are not based on reasons in the sense that Clifford had in mind, where to be based on reasons is to be *inferred* from reasons. This is the starting point for most contemporary discussions of self-knowledge. (See the essays in (Wright et al., 1998), especially, (Wright, 1998). See also (Moran, 2001)). But some theorists have thought that our beliefs about our mental states might still be based on reasons without being inferred from reasons. If we are clear about what is meant by being based on reasons for believing here, then that view is false, as I have argued. If what we mean is that our beliefs about our own mental states are rational only if they are psychologically based on reasons for believing in the rationaliz-

ing sense then this may very well be true. But theorists are often not careful to distinguish the two, so inferentialism has continued to look like a plausible option.

Once we have seen that inferentialism in all its forms is implausible in the case of introspection we might be encouraged to consider whether it is as plausible as it might seem in the case of beliefs about our immediate environment arrived at by sense perception, and in the case of basic a priori beliefs. I suspect that if we make all of the relevant distinctions in these contexts we will see that inferentialism is implausible there too.

CHAPTER 3

ACTIONS, REASONS, AND BECAUSES

3.1 Introduction

It is now standard in the philosophy of action to distinguish between three central kinds of explanations of actions: reason explanations, psychological explanations, and merely causal explanations.¹ The following are examples of each kind of explanation, respectively:

- (1) John is going to the pub because Jane is there.
- (2) John is going to the pub because he believes that Jane is there.
- (3) John is going to the pub because he has a drinking problem.

Questions then emerge as to the relation between these kinds of explanations. Do reason ex-

¹Unfortunately, usage varies. In “Actions, Reasons, and Causes”, Davidson called what we are calling psychological explanation *rationalizations* (Davidson, 1963, p. 3). He briefly discusses what we are calling reason explanations (Davidson, 1963, p. 8) but otherwise does not discuss them at all. He uses ‘reason explanation’ for what we are calling psychological explanations in (Davidson, 1976). Jennifer Hornsby uses the expression ‘reason explanation’ to cover both what we are calling ‘reason explanations’ and ‘psychological explanations’ (Hornsby, 2008, p. 250). There does, however, seem to be a growing tacit agreement to use ‘reason explanation’ as we are doing, and contrast it with psychological explanation. This is the usage in (Alvarez, 2010). Alvarez calls what we are calling reason explanations ‘reason explanations proper’. Setiya calls what we are calling reason explanations ‘reasons-explanations’ (Setiya, 2003, p. 349; Setiya, 2007, p. 23; Setiya, 2009b, p. 131). So too does Jonathan Dancy (Dancy, 2000). Hempel and Dray use the expression ‘rational explanation’. But it isn’t clear whether they mean to refer to what we are calling reason explanations or to what we are calling psychological explanations, or to both. I suspect that in the case of Dray it is the former, and in the case of Hempel it is the latter (Hempel, 1961; Hempel, 1965, p. 469; Dray, 1960).

planations entail psychological explanations? Are reason explanations a kind of psychological explanation? Do psychological explanations entail merely causal explanations? Are psychological explanations a kind of causal explanation? Do reason explanations entail merely causal explanations? Are reason explanations a kind of causal explanation?

The orthodox position in the philosophy of action is that reason explanations entail psychological explanations and are a kind of psychological explanation—call this position *psychologism about reason explanations*—and that psychological explanations entail merely causal explanations and are a kind of causal explanation—call this position *causalism about psychological explanations*—and that reason explanations entail merely causal explanations, and are a kind of causal explanation—call this position *causalism about reason explanations*.

I think that the orthodox position is right that reason explanations entail psychological explanations and that they are a kind of psychological explanation. I also think that the orthodox position is right that both psychological explanations and reason explanations entail merely causal explanations. (See (Dancy, 2000) for an argument against the orthodox view). The aim of this essay is to provide a new argument for psychologism about reason explanations. I argue on broadly linguistic grounds that reason explanations entail psychological explanations and that they are a kind of psychological explanation. The argument will be an extended inference to the best explanation of certain facts about the semantics and pragmatics of sentences like (1), the most important of which will be introduced in the next section. The remaining sections examine various potential explanations in detail. A more detailed overview of the argument will be given at the end of the next section.

Although the focus of this essay is the argument for psychologism about reason explanations, the conclusion has obvious implications for arguments for causalism about reason explanations. If reason explanations are a kind of psychological explanation, and psychological explanations are a kind of causal explanation, then reason explanations are a kind of causal explanation, and causalism about reason explanations is true.

3.2 The Language of Reason Explanations

Forget about reason explanations, psychological explanations, and causal explanations for a moment, and consider the following sentences:

- (1) John is going to the pub because Jane is there.
- (2) John is going to the pub because he believes that Jane is there.

Now, (1) seems to have a reading on which, at least relative to some contexts of use, it cannot be true unless John is going to the pub because he believes that Jane is there—that is, (1) seems to have a reading on which, at least relative to some contexts of use, it entails (2). But, (1) also has a reading—not necessarily a distinct reading—on which it does *not* entail that John is going to the pub because he believes that Jane is there. That is, on the face of it, (1) has a reading on which it can be true even though John is not going to the pub because he believes that Jane is there. On this reading, the meaning of (1) is analogous to the meaning of (3) above.

The main reason for thinking that (1) has these readings comes from the fact that speakers seem to be able to *use* (1) in particular contexts, to *convey* that John is going to the pub because he believes that Jane is there, and to use it in other contexts in a way that does not convey that John is going to the pub because he believes that Jane is there. Perhaps the simplest explanation of how they are able to *convey* that John is going to the pub because he believes that Jane is there, is that they can use (1) to *assert* something which entails that John is going to the pub because he believes that Jane is there. And perhaps the most simple explanation of how they are able to do *this* is that, at least relative to the context in which it is used, (1) is true if and only if things are as they are asserted to be. So, (1) seems to have a reading on which, at least relative to some contexts of use, it entails (2), and a reading—not necessarily a distinct reading—on which it does not.

One question that immediately arises is whether (1) really does have the readings it seems to have. On the one hand, there are various *semantically deflationary* explanations we might offer, which deny that (1) has a reading on which, at least relative to some contexts of use, it entails (2), and a reading on which it does not, and attempt to explain how (1) can be

used to *convey* (2) in some alternative way. Such explanations are semantically deflationary, since they do not posit distinct context-independent meanings for (1), and do not posit a context-dependent meaning for (1) on which its meaning is context-dependent in ways that go beyond the context-dependence which is ordinary acknowledge for sentences like (1). Such explanations would explain away our semantic intuitions.

Alternatively, perhaps we are confusing the conditions under which (1) would be true, relative to some context of use, with what (1) can be used to *imply*. Perhaps a sentence like (1) can be used to imply that John believes that Jane is at the pub in a particular context, even though (1) does not entail (2) relative to that context. Call this the *conversational implicature hypothesis*.

Or perhaps we are confusing the conditions under which (1) would be true, relative to some context of use, with an aspect of the conventional meaning of (1) which does not play a role in determining its meaning relative to some context of use. Perhaps a sentence like (1) can be used to conventionally imply that John is going to the pub because he believes that Jane is there, even though (1) does not entail (2) relative to any context of use. Call this the *conventional implicature hypothesis*.

Or perhaps we are confusing the conditions under which (1) would be true, relative to some context of use, with what a speaker must presuppose in order to use (1) to assert that John is going to the pub because Jane is there. Perhaps in particular contexts of use, a speaker must presuppose that John is going to the pub because he believes that Jane is there, in order to use (1) to assert that John is going to the pub, even though (1) doesn't entail (2) relative to that context of use. Call this the *pragmatic presupposition hypothesis*.

Or perhaps we are confusing the conditions under which (1) would be true, relative to some context of use, with what (1) can be used to *say*. Perhaps sentences like (1) can be used to say something which entails (2) in a particular context, even though (1) does not entail (2) relative to that context. Call this the *semantic/pragmatic underdetermination hypothesis*.

On the other hand, there are various *semantically inflationary* explanations we might offer, which hold that (1) has a reading on which, at least relative to some contexts of use, it entails (2), and a reading on which it does not.

Perhaps the most obvious explanation of this kind holds that ‘because’ has two different meanings, one on which (1) entails (2), and one on which it does not. This view is implicit in talk of the ‘because’ of reason explanations, as opposed to the ‘because’ of merely causal explanation. (Although such talk might be construed as not being committed on the question of whether ‘because’ has a single context dependent meaning, or has two different meanings). On this view, ‘because’ is lexically ambiguous, and (1) is derivatively lexically ambiguous. Call this the *lexical ambiguity hypothesis*. This is a semantically *inflationary* explanation, because it posits at least two distinct meanings for the word ‘because.’

Perhaps, alternatively, (1) itself has two different meanings, one on which it entails (2) and one on which it does not. One explanation of this kind holds that (1) has a reading on which it is syntactically elliptical for (2), and on which it therefore entails (2), and a reading on which it is not syntactically elliptical for (2), and on which it does not entail (2). Call this the *syntactic ambiguity hypothesis*. This hypothesis is semantically inflationary because it posits at least two distinct meanings for (1).

Perhaps, finally, (1) has a single context-independent meaning, but has a context-dependent meaning relative to some contexts, on which it entails (2) and a context-dependent meaning relative to other contexts, on which it does not entail (2). Call this the *context-dependence hypothesis*. This view avoids associating (1) with two distinct context-independent meanings, but is nonetheless semantically inflationary, since it posits a kind of context dependence which goes beyond what is typically acknowledged for explanatory language. It is typically held that explanatory language is semantically sensitive to salient alternatives to the explanans and explanandum (Dretske, 1972; van Fraassen, 1980; Achinstein, 1983).

3.2.1 *Explanation and Explanatory Language*

These are all hypotheses about explanatory language, about sentences like (1) and (2). But they are closely related to hypotheses about explanation itself, since (1) is the canonical expression of a reason explanation, and (2) is the canonical expression of a psychological explanation. It is reasonable to assume that the reason explanation expressed by (1) entails the psychological explanation expressed by (2) if and only if (1) entails (2). In what follows

we will look carefully at arguments for and against thinking that (1) has a reading on which it entails (2) and a reading on which it does not. I am going to argue that it does. And given this assumption, this will have immediate payoffs for the claim that reason explanations entail psychological explanations.

But we stand to learn more than just that reason explanations entail psychological explanation by examining sentences like (1) and (2), for different hypotheses bear differently on the question of whether reason explanations are a *kind* of psychological explanation, and on what kind of psychological explanation they are, if they are.

I am going to argue that the best explanation of the fact that (1) entails (2) is that ‘because’ is ambiguous, and that the ‘because’ of reason explanations means, roughly, ‘because *x* knows that’, where ‘because’ here is the ‘because’ of psychological explanation, and ‘*x*’ is an argument place for a subject of the main clause. Since the ‘because’ of reason explanations is analysed in terms of the ‘because’ of psychological explanation, it follows that (1) has an analysis in terms of (2). It is reasonable to assume that if (1) has an analysis in terms of (2), then the fact that John is going to the party because Jane is there, is partly constituted by the fact that John is going to the party because he believes that Jane is there, where the first ‘because’ is the ‘because’ of reason explanation, and the second ‘because’ is the ‘because’ of psychological explanation. Moreover, given the particular analysis of the ‘because’ of reason explanations in terms of the ‘because’ of psychological explanation, it turns out that reason explanations are a kind of psychological explanation.

3.2.2 Overview

What follows divides naturally into two parts. In the first part I discuss the semantically deflationary hypotheses. All of these hypotheses deny that (1) has a reading on which, relative to some contexts of use, it entails (2), and a reading—not necessarily a distinct reading—on which it does not. They all offer alternative explanations of the *fact*, and it is a fact, that (1) can be used to *convey* that John is going to the pub because he believes that Jane is there. In the second part, I discuss the semantically inflationary hypotheses. All of these hold that (1) has a reading on which, relative to some contexts of use, it entails (2), and a reading—not

necessarily a distinct reading—on which it does not. But they each offer alternative, more specific potential explanations.

The overall structure of the argument to follow is that of an inference to the best explanation. This being so, it will be necessary to consider *all* of the relevant alternative explanations of the relevant facts. We will not be in a position to determine which explanation is the *best* explanation, until a wide range of considerations for and against each hypothesis have been considered.

The amount of time I will spend discussing each hypothesis is proportionate both to how plausible I think the hypothesis is and how plausible I think that others will find the hypothesis. So I spend quite some time discussing the conversational implicature hypothesis, both because it seems to be the most plausible semantically deflationary hypothesis, and because I think that others will think that it is the most plausible semantically deflationary hypothesis, if not the most plausible hypothesis, deflationary or not. I spend quite some time discussing the lexical ambiguity hypothesis because I think that it is the most plausible hypothesis, and argue to that end. Many of the other hypotheses are treated here for completeness, and because they are likely to be raised as alternative explanations, even though I do not think they are particularly plausible.

Here is an overview. The next four sections examine the four semantically deflationary hypotheses. The most plausible of these, the conversational implicature hypothesis, turns out to be less plausible than it may seem at first glance. Various predictions of the hypothesis fail to be borne out, and it is difficult, if not impossible, to explain how a speaker could use (1) to merely imply (2) on the assumption that (1) has the kind of meaning semantically deflationary hypotheses attribute to it. The pragmatic presupposition hypothesis, the conventional implicature hypothesis, and the semantic/pragmatic underdetermination hypothesis, fare no better. Various predictions of these hypothesis fail to be borne out.

The following three sections examine the three semantically inflationary hypotheses. I begin with the syntactic ambiguity hypothesis. I argue that there is no plausible syntactic mechanism by which (1) could have a reading on which it is syntactically elliptical for (2). I then examine the lexical ambiguity hypothesis. I introduce some data which shows that, on

readings of sentences like (1) on which they entail (2), the relevant ‘because’-clause behaves like a *subject-oriented adverbial*—that is, it seems to bear a semantic relation to a subject of the main clause of which it is a constituent. Since the relevant ‘because’-clauses in ‘because’-sentences which do not have this reading do not behave this way, I conclude that ‘because’ is ambiguous—it has a meaning on which it bears a semantic relation to a subject of the main clause, and it is on a reading where ‘because’ has this meaning that (1) entails (2). Finally, I consider the context-dependence hypothesis as an alternative explanation of the opening data and the data bearing on whether ‘because’-clauses are subject-oriented adverbials. I argue that various predictions of the context-dependence hypothesis fail to be borne out, and that there are serious problems with attempting to explain how ‘because’-sentences could be context-dependent in the way needed to explain the opening data.

I am going to argue that the best explanation of the fact that (1) entails (2) is that ‘because’ is ambiguous, and that the ‘because’ of reason explanations means, roughly, ‘because *x* knows that’, where ‘because’ here is the ‘because’ of psychological explanation, and ‘*x*’ is an argument place for a subject of the main clause. Grice famously wrote that senses should not be multiplied beyond necessity (Grice, 1989, p. 47). And Kripke offers the following counsel: “Do not posit an ambiguity unless you are really forced to, unless there are really compelling theoretical or intuitive grounds to suppose that the ambiguity is really present” (Kripke, 2011, p. 118). I think it is sometimes assumed, on the basis of such remarks, that there is something problematic about positing an ambiguity. But of course, the truth in Grice and Kripke’s claim is merely that if one is going to posit an ambiguity, it had better be the *best* explanation of the linguistic facts. As they both pointed out, sometimes the best explanation of certain linguistic facts is not that some expression is ambiguous, but rather that it can be used in different ways. So my aim will be to show that the ambiguity hypothesis is the *explanation* of the facts.

3.3 The Conversational Implicature Hypothesis

This section explicates and evaluates the *conversational implicature hypothesis*.

There is a familiar pre-theoretical distinction between what someone says by uttering a

particular sentence and what someone implies by uttering that sentence. It is possible to say one thing, and mean something other than what one says, as when one mistakenly says something that one didn't mean to say—e.g., when one says 'anthropomorphic climate change' rather than 'anthropocentric climate change'—or when one says one thing and means the opposite—e.g. when one says that Amsterdam has great weather. Moreover, it is possible to say something, and to imply something in addition to what one says, as when one answers a question implicitly by explicitly providing an explanation of why the implicit answer obtains—e.g., when one says that John is ill in answer to the question of whether he is at the pub. We can draw these distinctions using ordinary verbs like 'say', 'mean', and 'imply', and the important distinctions are between things one means to say, things one does not mean to say, things one means to imply, and things one does not mean to imply. Following Grice, I will sometimes use the term 'implicate', rather than 'imply', to avoid confusion about what sentences imply—which is a matter of what the sentence means—and what speakers mean to imply by uttering a sentence—which is a matter of what speakers mean. I will use the term 'implicature' for *what* someone implies, and the term 'entailment' for *what* a sentence implies (Grice, 1989, p. 24).

How can I imply something more than what I say, or imply something other than what I say? I can do so by saying something in a particular conversational context, in a context where you assume that I am being cooperative, and therefore can reasonably expect me only to say things which I reasonably believe to be true, to say no more and no less than I reasonably believe to be true, to only say things which are relevant, and to use the most efficient and effective means to do so (Grice, 1989, pp. 26–27). I will successfully communicate in such a context, if you can figure out what I mean to communicate. You can figure out what I mean to communicate by working out what I mean to say or what I mean to imply. For example, if you ask me whether John is at the pub, I might say that he is ill. I say that he is ill, and imply that he is not at the pub. I mean to imply that he is not at the pub. You expect me to only say things which are relevant. I say that John is ill. You work out that I mean to say that he is ill, and to imply that he isn't at the pub. I imply this by conforming to all of your reasonable expectations. But sometimes I might imply something by manifestly failing to

conform to your expectations. If I say that John is someone who is unfamiliar with the taste of beer, I might imply that John is someone who is familiar with the taste of beer. I mean to imply that he is familiar with the taste of beer. I say something I think you will realise I do not reasonably believe. You work out that I mean the opposite of what I said.

Some expressions are used to imply the same thing in many conversational contexts, whereas other expressions are used to imply something only in very particular conversational contexts. The sentence ‘John is ill’ isn’t usually used to imply that John isn’t at the pub. It can be used to imply this in a context where someone has just asked whether John is at the pub. In contrast, the sentence ‘John drank some of Jane’s beer’ would be used to imply that John did not drink all of Jane’s beer. While nobody is tempted to think that ‘John is ill’ entails that John is not at the pub on the basis of the fact that it can be used to imply that John is not at the pub, someone might be tempted to think that ‘John drank some of Jane’s beer’ entails that John did not drink all of Jane’s beer on the basis of the fact that the sentence ‘John drank some of Jane’s beer’ would be used to convey that John did not drink all of Jane’s beer in many conversational contexts. The distinction here is between *particularized implicatures*—roughly, things that someone implies by using a particular sentence in very particular conversational contexts—and *generalized implicatures*—roughly, things that someone would imply by using a particular sentence in many conversational contexts (Grice, 1989; Levinson, 2000).

The sentence ‘John is going to the pub because Jane is there’ would be used to convey that John is going to the pub because he believes that Jane is there in many conversational contexts. It doesn’t seem to require a particular conversational context in order to be used to convey this. I think there would be little temptation to say that non-psychological ‘because’ sentences entail psychological explanations if the psychological explanations were particularized implicatures. Our claim is not that someone *can* use non-psychological ‘because’ sentences to (merely) imply a corresponding psychological explanation. The claim is, rather, that someone who utters a sentence like ‘John is going to the pub because Jane is there’ would, in an ordinary conversational context, reasonably be taken to be implying that John is going to the pub because he believes that Jane is there or saying something which entails this.

Someone who utters such a sentence would reasonably be taken to be saying or implicating this only if someone who utters such a sentence would, in ordinary conversational contexts, be implicating that John is going to the pub because he believes that Jane is there or saying something which entails this.

According to the generalized conversational implicature hypothesis, then, the best explanation of the fact that someone who utters a sentence like ‘John is going to the pub because Jane is there’ would, in an ordinary conversational context, convey that John is going to the pub because he *believes* that Jane is there is that the psychological explanation that John is going to the pub because he believes that Jane is there is a generalized implicature, and not an entailment—someone who utters the sentence, would, in an ordinary context, merely mean to imply the relevant psychological explanation; they would not say something which entails the relevant psychological explanation.

3.3.1 Cancelability and Reinforceability

We do not always mean to imply what we may reasonably be taken to imply to by saying what we do. Sometimes it might not be clear to an addressee that we do not mean to imply what we might reasonably be taken to imply. Sometimes it might not be clear to an addressee that we mean to imply what we mean to imply.

This has two important consequences. If you ask whether John is at the pub, and I say that he is fixing his bike, it will be reasonable for you to assume that I mean to imply that John is not at the pub. If I don’t mean to imply this, I can make it clear to you that I do not mean to imply it. I might say ‘He is fixing his bike. I mean, he told me he was fixing his bike. He might be doing it at the pub’. If I do mean to imply it, but suspect that it is not clear to you that I mean to imply it, I can make it clear to you by saying, ‘He is fixing his bike. He is not at the pub’.

These are important consequences, because one of the most important contrasts between what someone says and what someone merely implies is that saying something inconsistent with what you merely imply simply makes it clear what you mean, whereas saying something inconsistent with what you say, is to say inconsistent things, and that saying something

which entails what you merely imply simply makes it clear what you mean, whereas saying something which entails what you say, is to say something redundant (Grice, 1989, p. 44; Bach, 2001, p. 18; Levinson, 2000, p. 15; Potts, 2007, p. 670).

These consequences give rise to two standard linguistic tests for conversational implicatures and ordinary entailment. These tests involve judgements about inconsistency and redundancy. We can ask whether something someone says is inconsistent with something else they say, or whether something someone says is redundant, given something else they have said. Judgements about inconsistency and redundancy are standardly taken to track relations between what is said, and not between what is merely implied. Indeed, this is because if you say something inconsistent with something you merely imply, then you *cancel* the implication, and if you say something redundant, given something you have implied, then you *reinforce* the implication. Either way, you just make clear what you mean.

Let us ask, then, whether someone who utters the sentence ‘John is going to the pub because Jane is there’, in a conversational context where it would be reasonable to take them to be conveying that John is going to the pub because he believes that Jane is there, can make it clear that this is not what they mean, without inconsistency. To my ears, they cannot. Someone who says ‘John is going to the pub because Jane is there, but he isn’t going to the pub because he believes that Jane is there’, in such a context, seems to say something inconsistent.

It is important that we do not misconstrue this test. To say that an implicature can be cancelled is not the same thing as saying that there are contexts in which the sentence can be used to say something which does not entail the purported implicature. There are contexts in which the sentence ‘I am Australian’ can be used to say something which does not entail that the author of this essay is Australian. That doesn’t show that the proposition that the author of this essay is Australian is a mere implicature. Indeed, in contexts where an utterance of this sentence would reasonably be taken to convey that the author of this essay is Australian, it is not possible to say ‘But the author of this essay isn’t Australian’ without contradiction or inconsistency. The mere possibility of asserting the conjunction or the crucial statement and the negation of the suspect condition in some context does not tell us anything about whether

the suspect condition is an implicature in some other context.

Unfortunately, the application of the test for cancelability is not so straightforward. It is difficult to distinguish, using this test alone, whether the follow up ‘but he isn’t going to the pub because he believes that Jane is there’, is intended to make clear what someone means to *say*, by disambiguating the first part of the sentence, or whether it is intended to make clear that one does not mean to *imply* what one might otherwise be taken to mean to imply. We would do well to remind ourselves at this point that cancelability is taken to be a necessary feature of implicatures, but not a sufficient feature. Still, the fact that there are conversational contexts in which someone who says ‘John is going to the pub because Jane is there, but he isn’t going to the pub because he believes that Jane is there’, seems to say something inconsistent, suggests that the corresponding psychological explanation is not *always* merely implicated. Indeed, it suggests that the corresponding psychological explanation is something which is *entailed* by what the speaker says.

We can now turn to a related test. If someone means something other than, or more than, what they say, then they can make this clear without redundancy. Let us ask, then, whether someone who utters the sentence ‘John is going to the pub because Jane is there’, in a conversational context where it would be reasonable to take them to be conveying that John is going to the pub because he believes that Jane is there, can make it clear that this is part of what they mean, without redundancy. To my ears, they cannot. Someone who says ‘John is going to the pub because Jane is there. Indeed, he is going to the pub because he *believes* that Jane is there’ in such contexts, says something redundant by uttering the second sentence. Again, care must be taken with this test. But here it is easier to take that care. Again, it is irrelevant that someone can say these things in a particular conversational context without saying something redundant by uttering the second sentence. The question is whether they can do so in contexts where they would reasonably be taken to be conveying the corresponding psychological explanation in uttering the first sentence. In such conversational contexts, the second thing they say is clearly redundant. This strongly suggests that the suspect condition is not a mere implicature, but is something which is entailed by what the speaker says, and is thus entailed by the sentence itself relative to that context.

On the face of it then, the corresponding psychological explanation seems to be neither cancellable nor reinforceable. This strongly suggests that it is not a mere implicature—something implicated by someone who utters the relevant sentence in ordinary contexts. I don't mean to suggest that these considerations are decisive. Our judgements here are subtle and fallible. My argument isn't an argument by elimination of alternatives. It is an argument to the best explanation. As such, we are going to have to weigh an array of reasons for and against particular candidate explanations. These reasons weigh against this candidate explanation. But if all the other explanations are even worse, then the generalised conversational implicature hypothesis may turn out to be the best, and perhaps even good enough for us to reasonably infer that it is true.

3.3.2 Calculability

So far we have largely been on the defensive, although we have made a small offensive move. We have been on the defensive, because theorists often treat features of implicatures as being sufficient for implicatures, rather than merely necessary. I have suggested that in the case of reinforceability and calculability, our examples might fail to meet the necessary condition. The main challenge for the generalized conversational implicature hypothesis is that of explaining *how* it is that speakers can implicate what they are supposed to implicate by uttering the sentence 'John is going to the pub because Jane is there'. It is widely accepted that what one can use a sentence to implicate often depends, at least in part, on what the sentence means. So the question is how someone can implicate what they are supposed to implicate by uttering the sentence 'John is going to the pub because Jane is there' (Grice, 1989, p. 31).

Before turning to such explanations it will be helpful to have candidate meanings for (1) on the table which are semantically deflationary in the sense that they do not posit an ambiguity or more context sensitivity than is standardly recognised for 'because' sentences. There are two salient hypotheses about the meaning of (1) which a proponent of our semantically deflationary hypotheses might adopt. On one hypothesis (1) expresses a merely causal explanation, like (3) does. On another hypothesis, (1) is general between reason explanations,

psychological explanations, and causal explanations. To understand this view, notice that we would recognise a distinction between the kind of explanations expressed by ‘Vixy is a vixen because Vixy is a female fox’ and ‘Vixy is dead because she was hit by a car’. The former is what we might call a metaphysical explanation, and the latter is a merely causal explanation. But it is very plausible that ‘because’ is semantically general with respect to the distinction: it doesn’t distinguish between them. On either of these views, if someone utters the sentence ‘John is going to the pub because Jane is there’, they say something which does not entail that John is going to the pub because he knows, or at least believes, that Jane is there. This is secured by the fact that on these views ‘because’ is semantically general between (1) and (3), and one calculates entailments by generalizing to the worst case. Since the worst case is one where (1) is true because of a merely causal explanation, which doesn’t entail (2), it follows that (1) doesn’t entail (2).

Recall the point that we can imply things against a particular assumption of cooperation and expectations which follow from that assumption. I can use the sentence ‘John is ill’ to imply that he isn’t at the pub, because you expect me to say something which is relevant. Since we were discussing the question of whether John is at the pub, I can imply that he isn’t at the pub by saying that he is ill. We have a pretty good impression of *how* I managed to imply this by saying that John is ill. Similarly when I say that John is unfamiliar with the taste of beer and imply that he is familiar with the taste of beer. We have a pretty good idea of *how* I managed to do this. I said something which was manifestly false, and so is something that I do not reasonably believe. But since you still assumed I was being cooperative, you worked out that I meant the exact opposite of what I said.

Here is one potential explanation of the facts, which draws on the first hypothesis about the meaning of our sentence. I can implicate that John is going to the pub because he believes that Jane is there by uttering a sentence which expresses a causal explanation with the fact that Jane is at the pub as its explanans. Although by doing so I say that the fact that Jane is at the pub causally explains the fact that John is going to the pub, I do not mean what I say. I might assume that this will be manifest to you. So you will realise that this is not what I mean to say. So you will figure that I must mean something else. The most obvious thing I

could mean is that John is going to the pub because he *believes* that Jane is there. I assume that you can work this out. If all goes well, I will manage to imply that John is going to the pub because he believes that Jane is there by uttering the sentence ‘John is going to the pub because Jane is there’.

There are several problems with this explanation. The first is that we typically *do* mean what we say when we say that John is going to the pub because Jane is there. So we do not say one thing, and mean something quite different. The second is that addressees typically assume that we mean what we say when we say such things. When someone utters the sentence ‘John is going to the pub because Jane is there’ we do not feel that they mean something other than what they say: ‘Well what Jack *said* was that John is going to the pub because Jane is there, but what he *meant* is that he is going to the pub because he believes that Jane is there.’

If we cannot explain how a speaker might imply the relevant psychological explanation in terms of their *flouting* an addressee’s expectations, then perhaps we can explain how a speaker might imply the relevant psychological explanation by *conforming* to an addressee’s expectations. The model here is given by our earlier example of someone who implies that John is not at the pub by asserting that he is ill. Might a speaker imply that John is going to the pub because he believes that Jane is there in something like the way a speaker might imply that John is not at the pub when they say that he is ill?

The most plausible way of filling out this explanation begins with the idea that the conversational context is one in which the question under discussion is what psychologically explains the fact that John is going to the pub. In such a context, an addressee will expect the speaker to be relevant, and to speak to the question under discussion. We saw that when the question under discussion is whether John is at the pub, someone can answer this question by saying that John is ill and implying that he isn’t at the pub. Such an answer is clearly relevant, and it provides additional information. Our question, then, is whether someone can use the sentence ‘Jane is going to the pub because Jane is there’ to imply the corresponding psychological explanation that John is going to the pub because he believes that Jane is there, if the sentence does not entail that John is going to the pub because he believes that Jane is there. Obviously, if the sentence ‘Jane is going to the pub because Jane is there’ did entail the

relevant psychological explanation, one could provide a relevant answer by saying that John is going to the pub because Jane is there. The answer is relevant, because it entails an answer to the question under discussion, and it might be preferable to an explanation in terms of John's beliefs, since such an explanation might imply that Jane is not at the pub. (This point about the pragmatics of psychological explanations and reason explanation is often noted. See: (Smith, 2004; Alvarez, 2010, Ch. 5)).

Consider the first semantic hypothesis. When someone says that John is going to the pub because Jane is there, they say merely that the latter causally explains the former. So they say something which does not entail that John is going to the pub because he believes that Jane is there. The thought here is that one might give a relevant explanation by giving the cause of John's psychological state, which, in turn, is implied to psychologically explain his going to the pub. The model here would be someone who answers the question 'Why did the bridge collapse?' by saying that John planted a bomb on it, and implying that the bridge collapsed because the bomb exploded. This explanation has some initial plausibility. But to see that it cannot be correct, consider the sentence 'John is going to the pub because Jane will be there later'. It is doubtful that the fact that Jane will be at the pub later causally explains the fact that John is going to the pub now. In fact, it is generally quite implausible to assume that there is an explanatory connection between one's belief and what is believed. This suggests that this model is inadequate.

Consider the second semantic hypothesis, which can at least allow that 'John is going to the pub because Jane will be there later' to be true, since it is general between reason explanations and causal explanations. It is difficult to see how one might use a sentence which is general in this way to imply a particular psychological explanation. If the question under discussion is what the psychological explanation of John's going to the pub is, and I offer you an answer which is general between a causal explanation in terms of the fact that Jane is there and a reason explanation in terms of the fact that Jane is there, then, even if reason explanations entail psychological explanations, I am being needlessly irrelevant. And this is due to the generality of what I say. An analogy will help. The term 'sibling' is general between male and female sibling. Suppose that it is well known that John has one male

sibling, Jack, and one female sibling, Jill. Suppose the question under discussion is whether Jack was at the pub. Let's ask whether someone could imply that Jack was at the pub by asserting that one of John's siblings was at the pub. Even in a context in which it was known that Jill was not at the pub, it is hard to see how one could do so. One would be needlessly irrelevant. And it is due to the generality of one's answer that one is so. Similarly, a reason explanation would be relevant, but I do not assert it when I assert that John is going to the pub because Jane is there. It is precisely because this semantic hypothesis tries to avoid the entailment that the explanation doesn't work here.

3.3.3 Embedded Contexts

So far we have been trying to tease apart what is said and what is implied by someone who utters a declarative sentence. Our ultimate aim has been to argue that someone who says that John is going to the pub because Jane is there in an ordinary conversational context, says something which entails that John is going to the pub because he believes that Jane is there. From this we want to infer that (1) entails (2), at least relative to certain contexts of use. But there is a more direct way of arriving at such semantic conclusions, and that is by considering (1) in embedded contexts. So far we have been considering utterances of (1) in which it occurs unembedded. By considering utterances of (1) where it occurs embedded in a larger expression, we might be better able to get at its meaning (Grice, 1989, pp. 45-46).

There are several embedded contexts of interest. The first is that of the antecedent of a conditional. Consider the sentence 'Bill is meeting a woman this evening'. The conditional 'If Bill is meeting a woman this evening, then Bill is meeting a woman other than his wife' will be considered false. This suggests that the proposition that Bill is meeting a woman other than his wife is something which is merely implicated by someone who says that Bill is meeting a woman this evening. Now consider: 'If John is going to the pub because Jane is there, then John is going to the pub because he believes that Jane is there'. This is typically acceptable, and so true, in contexts in which someone who utters the antecedent would reasonably be taken to be conveying the antecedent. Again, it is irrelevant that there are contexts in which the conditional is false. This doesn't show that there are no contexts in which the antecedent

does not entail the consequent or that a speaker who utters the antecedent must presuppose the psychological explanation. (Importantly, this test doesn't distinguish between entailments and presuppositions. Further tests are needed. But it tells pretty strongly against the generalized conversational implicature hypothesis).

Another embedded context is under 'suppose'. If someone says 'Suppose that John is going to the pub because Jane is there', they will typically be taken to be inviting you to suppose something which entails that John is going to the pub because he believes that Jane is there. In supposing what you have been asked to suppose, you will suppose that John knows, or at least believes, that Jane is at the pub, and that this is why John is going to the pub. Since this sentence can be used to invite someone to suppose this, this suggests that the embedded sentence entails that John is going to the pub because he believes that Jane is there. Similarly for another embedded context. It seems that someone can use 'It is not the case that John is going to the pub because Jane is there' to deny that John is going to the pub because he believes that Jane is there. Indeed, consider someone who says: John doesn't even believe that Jane is at the pub, so it is not the case (or it cannot be the case) that he is going to the pub because Jane is there.' This would make sense only on the hypothesis that we are dealing with an entailment rather than an implicature here.

We have seen in this section that a tempting explanation of our opening facts in terms of conversational implicature is not particularly plausible. Such an explanation is tempting because it is semantically deflationary. But, as we have seen, it is because it is semantically deflationary, that it cannot explain what it would have to explain. Let us now turn to alternative semantically deflationary hypotheses.

3.4 The Pragmatic Presupposition Hypothesis

It is sometimes said that someone who asserts, say, that John is going to the pub because Jane is there *presupposes*, among other things, that John is going to the pub because he believes that Jane is there. Again, there is a pre-theoretical distinction between what a speaker says or implies, and what a speaker presupposes. If someone who asserts that John is going to the pub because Jane is there presupposes a particular psychological explanation, then this might

provide an alternative explanation of why the sentence ‘John is going to the pub because Jane is there’ is often used to convey that John is going to the pub because he believes that Jane is there. The explanation can proceed as follows. Although someone who asserts that John is going to the pub because Jane is there does not *assert* or *imply* that John is going to the pub because he believes that Jane is there, they may mean to *convey* or simply *convey* without meaning to convey, that John is going to the pub because he believes that Jane is there. They may do so, because an addressee who does not already know that John is going to the pub because he believes that Jane is there may work out that the speaker is presupposing that this is so in saying what she is saying and come to believe it on the basis of this (Stalnaker, 1999b, p. 39; Stalnaker, 1999a, pp. 51–52). In this way, it might be claimed, the sentence ‘John is going to the pub because Jane is there’ can often be used to convey that John is going to the pub because he believes that Jane is there. (As with implying, it is, in the first place speakers who presuppose things (Stalnaker, 1999a, p. 50)).

Much more needs to be said to fill out this hypothesis. First of all, this explanation would do nothing to show that the sentence does not *entail* that John is going to the pub because Jane is there. It is perfectly compatible with this hypothesis that the sentence does entail this. A speaker who says that John realised that Jane was not at the pub must presuppose that Jane was not at the pub. And so they might convey that Jane was not at the pub to an addressee who didn’t already know this. But the sentence ‘John realised that Jane was not at the pub’ entails that Jane was not at the pub. It cannot be true unless Jane was not at the pub (Stalnaker, 1999a, p. 54). Secondly, this explanation would do nothing to show that the sentence is not ambiguous between reason explanations and merely causal explanations. The best explanation of the fact that a speaker must presuppose that John is going to the pub because Jane is there might be that the sentence is a presupposition trigger—that is, it is part of the meaning of the sentence that a speaker must presuppose that something is so. In the case of ‘John quit smoking’, the best explanation of why the speaker must presuppose that John once smoked is that the sentence itself requires the speaker to in virtue of its meaning. That said, the question of whether a speaker must presuppose something in a particular conversational context, or across a range of such contexts, is independent of the question of what best ex-

plains this (Stalnaker, 1999a, p. 55). It may still turn out that the best explanation here is that ‘because’-sentences are presupposition triggers. (We can distinguish here between merely conversationally triggered presuppositions, which are heavily determined by what is already being presupposed in the conversational context, and conventional presuppositions which are triggered by the uttered sentence (Potts, 2004, p. 23)).

There is a model for presuppositions where the sentences involved do not also entail what is presupposed, and are not presupposition triggers. The model is given by conversational contexts in which a question with a particular presupposition has been asked, and the answer requires the speaker to make a particular presupposition in answering. John is going to a pub. But there are several pubs in town. Where is John? you ask. He is at Murphy’s, I answer. I neither say that Murphy’s is a pub, nor mean to imply that Murphy’s is a pub. It is not even clear that I mean to convey this. (Of course, if I knew that you were not aware that Murphy’s was a pub, I might mean to convey it. But it may very well be that I accidentally convey it, because I do not realise that you are not aware that Murphy’s is a pub). Nonetheless, I may convey to you that Murphy’s is a pub. Question-answer pairs are a rich source of presuppositions which are not to be explained in semantic terms. Nobody would dream of hypothesising that someone who asserts that John is at Murphy’s must presuppose that Murphy’s is a pub in all contexts. How does this happen? I assert something which, in the context, requires me to presuppose that Murphy’s is a pub.

This model is particularly fitting, because the sentence ‘John is at Murphy’s’ does not entail that Murphy’s is a pub. Moreover, there is little temptation to claim that this sentence triggers the presupposition that Murphy’s is a pub, which is to make a claim about the context-independent meaning of the sentence. In short, the example is an example where a speaker must presuppose that something is so, but where (i) the sentence does not also entail what is presupposed by the speaker and (ii) the best explanation is purely pragmatic. It is due to particular features of the conversational context that the speaker must presuppose that Murphy’s is a pub.

The thought, then, is that, in the cases of interest to us, the relevant conversational contexts are such that they require the speaker to presuppose that John is going to the pub because he

believes that Jane is there—the speaker must believe that John is going to the pub because he believes that Jane is there, and assume that the addressee knows this, or will come to believe it on the basis of what she says. This is certainly an alternative explanation of the fact that such sentences are often used to *convey* particular psychological explanations, one which is compatible with the hypothesis that the sentences do not entail particular psychological explanations, and is compatible with the hypothesis that ‘because’-sentences are semantically general between reason explanations and psychological explanations.

There are further features of our analogy which suggest that this hypothesis cannot be right, however. In short the problem is this. Consider:

- (4) John is at Murphy’s.
 - a. John isn’t at Murphy’s.
 - b. Is John at Murphy’s?
 - c. John might be at Murphy’s.
 - d. If John is at Murphy’s, he is in trouble.

If we are in a conversational context where my assertion that John is at Murphy’s conveys that Murphy’s is a pub, then we are in a conversational context where my assertion that John isn’t at Murphy’s conveys this, my asking whether John is at Murphy’s conveys this, my asserting that if John is at Murphy’s, he is in trouble, conveys this, so too does my asserting that John might be at Murphy’s. These marks are missing in the case of our ‘because’-sentences. Consider:

- (5) John is going to the pub because Jane is there.
 - a. John is not going to the pub because Jane is there.
 - b. Is John going to the pub because Jane is there?
 - c. John might be going to the pub because Jane is there.
 - d. If John is going to the pub because Jane is there, then Jack will be annoyed.

As these examples demonstrate, if we are in a conversational context where my assertion that John is going to the pub because Jane is there conveys that John is going to the pub because he

believes that Jane is there, it does not follow that my denial, or my asking, or my expression of uncertainty, convey that John is going to the pub because he believes that Jane is there. (These examples do demonstrate, however, that these sentences require speakers to presuppose the explanandum. For an explanation of this in terms of an event semantics which assimilates it to a case of the presuppositional requirements on quantificational quantifier-restrictor-scope structure see: (Larson, 2004; Kawamura, 2008)).

3.5 The Conventional Implicature Hypothesis

An alternative hypothesis on which (1) does not have a reading on which it entails (2) and a reading on which it does not entail (2), holds that, on the reading on which it seems to entail (2), (1) merely conventionally implicates (2). What might encourage us in pursuing this hypothesis is that conventional implicatures, unlike conversational implicatures, cannot be cancelled, and cannot be reinforced, and they are not entailed by what is said, if what is said is tied closely to the meaning of the sentence uttered.

Grice famously introduced conventional implicatures only to go on and contrast them with conversational implicatures. For Grice, the important distinction was that conversational implicatures are generated by an interaction between the cooperative principle and the principles of rational communication, and are independent of the conventional meaning of the sentences involved, whereas conventional implicatures are tied to the conventional meanings of the sentences involved (Grice, 1989, pp. 44–45).

Conventional implicatures are a matter of some controversy, however. And some theorists have come to doubt that there is such a thing. (A related issue here is that some theorists use ‘presuppositions’ and ‘conventional implicatures’ interchangeably (Potts, 2004, pp. 11–13)). This is largely due to the fact that some standard examples of conventional implicatures turn out to be better treated as part of what is said. The paradigm example here is ‘but’. Consider the sentence ‘John is going to the pub but he will be home in time for dinner’. This suggests that John is not usually home in time for dinner when he goes to the pub. And it might be thought that this is something which is merely implied. But, there are good arguments for thinking that this is part of what is said by someone who utters this sentence (Bach, 1999), so

it is not a good example of a conventional implicature.

Christopher Potts has recently revived the category of conventional implicatures, however. Consider the following example (Potts, 2004, pp. 13–14):

- (6) Conventional implicatures, which are the topic of this section, are more controversial than conversational implicatures.

An utterance of this sentence will certainly convey that conventional implicatures are the topic of this section. But there is an intuitive sense in which this is not part of what I say. Rather, it seems to be a comment on something I am saying, a side remark of sorts (Potts, 2004, p. 43). This is an example of a supplement, or, more precisely, a parenthetical. Potts lists supplements (appositives, and parentheticals), speaker-oriented and topic-oriented adverbials, and expressives, as expressions associated with conventional implicatures (Potts, 2004, pp. 13–22).

It might be thought that ‘because’ sentences should be added to the list. Perhaps, (1) conventionally implicates (2). After all, conventional implicatures cannot be cancelled or reinforced. So, to this extent, they pattern with ordinary entailments, but not with conversational implicatures (Potts, 2004, p. 10). Consider:

- (7) Conventional implicatures, which are the topic of this section, are more controversial than conversational implicatures, but this section isn’t about conventional implicatures.

As this sentence shows, the purported implicature cannot be explicitly cancelled without contradiction. There are serious problems for this suggestion, however. One is that it is a mark of conventional implicatures that they persist in certain contexts (Potts, 2004, p. 31). Consider:

- (8) Jane believes that John, who is Jane’s brother, is going to the pub because Jane is there.

An utterance of (8) will convey that John is Jane’s brother. Compare:

- (9) Jack believes that John is going to the pub because Jane is there.

An utterance of (9) will not convey that John is going to the pub because he believes that Jane is there. So the suspect condition does not behave like a conventional implicature.

Moreover, conventional implicatures persist in other contexts too, contexts which are called presupposition holes (negation, interrogatives, modals, and conditionals) (Karttunen, 1973; Potts, 2004, p. 31). Consider the following paradigm:

- (10) John, who is Jane's brother, is going to the pub.
- a. John, who is Jane's brother, is not going to the pub.
 - b. Is John, who is Jane's brother, going to the pub?
 - c. John, who is Jane's brother, might be going to the pub.
 - d. If John, who is Jane's brother, is going to the pub, Jane must be there.

Compare:

- (11) John is going to the pub because Jane is there.
- a. John is not going to the pub because Jane is there.
 - b. Is John going to the pub because Jane is there?
 - c. John might be going to the pub because Jane is there.
 - d. If John is going to the pub because Jane is there, then Jack will be annoyed.

Whereas the suspect condition is conveyed in all of the former examples, it is not conveyed in the latter. Utterances of the latter do not convey that John is going to the pub because he believes that Jane is there. They do not even convey that John believes that Jane is there. They do not even convey that Jane is at the pub. At best, each conveys that John is going to the pub.

This suggests that we cannot account for the purported entailment in terms of conventional implicatures. Moreover, even if we could, since (1) can be used in a way which does *not* convey (2), it must be ambiguous, in the sense that it has one conventional meaning associated with a particular conventional implicature, and another conventional meaning associated with another conventional implicature. So, since this hypothesis requires (1) to be ambiguous, in the sense of 'meaning' associated with conventional implicature, it is not clear why one shouldn't just accept that (1) has two meanings, in the ordinary sense of 'meaning'.

3.6 The Semantic/Pragmatic Underdetermination Hypothesis

According to the final semantically deflationary hypothesis we will consider, sentences like (1) can be used to say something which entails (2) in a particular context, even though (1) does not entail (2) relative to that context. This hypothesis is somewhat more radical than the rest. We needn't deny that a sentence might be used to say something which entails something else, even though the sentence does not entail it, even relative to that context of use. Perhaps, as Kent Bach has argued, some declarative sentences are semantically incomplete, they do not have truth conditions, even relative to a context of use, and so do not entail anything (Bach, 1994). The question then is whether sentences like (1) are semantically incomplete in this sense. This is a hypothesis worth bearing in mind, but it is a radical hypothesis to adopt about (1). There is nothing to suggest that (1) is semantically incomplete.

One might hold an even more radical view and hold that even though a particular sentence has truth-conditions relative to some context of use, it can be used to say something which entails something the sentence itself does not entail. On this view, what a sentence can be used to say departs significantly from what it means. The main objection to this hypothesis is that embedded contexts reveal that it is not very plausible to posit such a gap in the case of (1). Indeed, most radical hypotheses of this kind hold that someone can use a particular sentence to say something which does not entail what the sentence itself entails relative to the context of use. Thus, some hold that someone can use 'Everyone is at the pub' to assert something which does not entail that everyone in the universe is at the pub—for instance, it might be used to assert that all of John's friends are at the pub—even though the sentence 'Everyone is at the pub' entails that everyone in the universe is at the pub, relative to the context in which it is used. The opposite would be needed here. I will set aside this radical alternative.

3.7 The Syntactic Ambiguity Hypothesis

We turn now to the first semantically inflationary hypothesis we will consider, namely, the syntactic ambiguity hypothesis. According to this hypothesis (1) has a reading on which it is

syntactically elliptical for (2), and it also has a reading on which it is not. More specifically, according to this hypothesis, ‘John is going to the pub because Jane is there’ has a reading on which it is syntactically elliptical for ‘John is going to the pub because he believes that Jane is there’ and has a reading on which it is not elliptical for this. On this view (1) is *syntactically ambiguous*. It has two different syntactic structures, and, the two different readings are a result of these two different syntactic structures. It is easy to see why (1) would have a reading on which it entails (2) on this view, for on one of the readings, it simply has the same meaning as (2).

To say that (1) has a reading on which it is syntactically elliptical for (2) is to say that it has a reading on which it is elliptical for (2) in the sense that ‘John is going to the pub, and Jane is too’ is elliptical for ‘John is going to the pub, and Jane is going to the pub too.’ The latter is a paradigm example of syntactic ellipsis. The thought is that ‘John is going to the pub, and Jane is too’ has the *same* syntactic structure as ‘John is going to the pub, and Jane is going to the pub too’, it is just that the verb phrase ‘going to the pub’ is not written or pronounced in the second conjunct of the former. This is called ‘syntactic’ ellipsis, because it is thought to be strictly governed by the rules of syntax. It contrasts with ‘pragmatic’ ellipsis, which is just another name for implicature. (When philosophers say that reason explanations are elliptical for psychological explanations, it is hard to know whether they mean that the former implicate the latter, or that the former are syntactically elliptical for the latter. See for example, the claim that reason explanations are elliptical for psychological explanations in (Cullity and Gaut, 1997, p. 2)).

One needs only a text-book level understanding of syntactic ellipsis to see that this explanation is implausible. There are two central problems. The first problem is the lack of a linguistic antecedent for the elided material. The second problem is that the explanation cannot be made out on any known model for syntactic ellipsis.

Let’s take the first problem first. Consider the following examples:

(12) John is going to the pub, and Jane is too.

(13) John knows that someone is going to the pub, but he doesn’t know who.

(14) John ordered a beer and Jane a cider.

These examples are examples of what linguists call, ‘verb phrase ellipsis’, ‘sluicing’, and ‘gapping’ respectively (Carnie, 2006; Freidin, 2012). These are the most common and best understood kinds of syntactic ellipsis. In each example there is a clear linguistic antecedent for the elided material. Most syntacticians believe that a clear linguistic antecedent is a necessary condition for syntactic ellipsis (Merchant, 2004; Culicover and Jackendoff, 2005; Carnie, 2006; Freidin, 2012). The linguistic antecedent need not be within the same sentence, however, but it must be present in the discourse. Moreover, ‘antecedent’ can be misleading. Consider: ‘Jane doesn’t know who, but she knows that someone went into the tunnel’. Here the elided material is present in the immediate discourse, although it isn’t in the antecedent discourse. The first problem, then for the syntactic ambiguity hypothesis is that there need not be an occurrence of ‘believed’ in the immediate discourse of ‘John is going to the pub because he has a drinking problem’ in order for it to entail that John is going to the pub because he knows that he has a drinking problem. So, the entailment we are interested in cannot be explained in terms of ‘John is going to the pub because he has a drinking problem’ being syntactically elliptical for ‘John is going to the pub because he knows that he has a drinking problem.’

The second problem for the syntactic ambiguity hypothesis is that the required syntactic explanation cannot be made out on any known model for syntactic ellipsis. Since sluicing is a phenomenon associated with ‘wh’ words, we can rule out sluicing as a possible explanation. That leaves verb phrase ellipsis and gapping as the two most salient possible explanations. (There are other kinds of ellipsis, but we will not consider them here as they are even less plausible candidates).

Consider another example of verb-phrase ellipsis: ‘Jane is going to Mark’s party, and Mary is too’. This is called verb-phrase ellipsis because the entire verb phrase ‘going to Mark’s party’ is elided in ‘Jane is going to Mark’s party, and Mary is too’. If we tried to explain the claim that a typical utterance of ‘Jane is seeking medical advice because she has a drinking problem’ is an utterance of a sentence with a syntactic structure identical to ‘Jane is seeking medical advice because (she knows that) she has a drinking problem’ in terms of verb phrase ellipsis we would expect the entire phrase ‘knows that she has a drinking problem’ to

be elided. But you cannot simply elide the verb and leave the complement of the verb in place. Moreover, verb phrase ellipsis cannot account for the fact that ‘she’ is elided, since ‘she’ is not part of a verb phrase. In short, the claim cannot be explained on the model of verb phrase ellipsis.

Now let’s consider gapping, as gapping does allow for the elision of single verbs. Consider the example again: ‘Jane bought a television, and John a radio’. Here the verb ‘bought’ is elided. So you might think that gapping can account for the elision of ‘knew’ in ‘Jane is seeking medical advice because (she knows that) she has a drinking problem’. But there are several problems with this proposal. The first is that gapping is not permitted in subordinate clauses. Consider: ‘Jane bought a television before John a radio’. In this sentence the complement of ‘before’ is a subordinate clause, and, as the fact that this sentence is not grammatical demonstrates, gapping is not permitted in subordinate clauses. The complement of ‘because’ in a because sentence is a subordinate clause. So gapping is not permitted. Moreover, as with the case of verb phrase ellipsis, we still do not have a mechanism for eliding ‘she’. In short, then, given these two problems, the claim that a typical utterance of ‘Jane is seeking medical advice because she has a drinking problem’ is an utterance of a sentence with a syntactic structure identical to ‘Jane is seeking medical advice because (she knows that) she has a drinking problem’ is not empirically viable.

3.8 The Lexical Ambiguity Hypothesis

Another potential explanation of the fact that (1) has the readings it seems to have is that ‘because’ has more than one sense or meaning, it is ambiguous, and so (1) is lexically ambiguous. An expression is *ambiguous* if it has more than one meaning (Gillon, 1990, p. 394). More precisely, an expression is ambiguous if it has more than one context-independent meaning. The different meanings of an expression are often referred to as different ‘senses’ of the expression (Cruse, 1986, p. 49). An expression is ambiguous, then, if it has more than one sense. The term ‘bat’ is ambiguous between nocturnal mammals and something designed to hit a ball. The simplest version of this explanation holds that ‘because’ has one sense when it is used in a sentence which expresses a psychological explanation or a merely causal ex-

planation, and it has another sense when it is used in a sentence which expresses a reason explanation, and that here it means, roughly, ‘because x believes that’ or ‘because x knows that’, where ‘because’ here is the ‘because’ of psychological explanation, and ‘x’ is an argument place for the subject of the main clause, of which the ‘because’-clause is a part. I will call this the *lexical ambiguity hypothesis*.

3.8.1 *Because Clauses as Subject Oriented Adverbials*

Let me start with the strongest argument for the ambiguity of ‘because’. Consider the following sentences:

(15) John examined Jane because her test results were troubling.

(16) Jane was examined by John because her test results were troubling.

Now (15) has the two readings we are familiar with, one which entails that John examined Jane because he knew that Jane’s test results were troubling, and one which does not. Strikingly, (16) has *three* readings. It has a reading on which it does not entail that either Jane or John knew that Jane’s test results were troubling. It has a reading on which it entails that *Jane* knew that her test results were troubling, and a reading on which it entails that *John* knew that Jane’s test results were troubling. More precisely, (16) has a reading on which it entails neither (17) nor (18). It has a reading on which it entails (17). And it has a reading on which it entails (18).

(17) Jane was examined by John because John knew her test results were troubling.

(18) Jane was examined by John because she knew her test results were troubling.

I suggested earlier that ‘because’ has one meaning, represented by ‘because_r’ on which it expresses a relation which obtains between one fact and another if and only if the former fact is rationally explained by the fact that someone knows the latter fact. By ‘someone’ I didn’t mean ‘at least one person’, I meant ‘someone in particular’. Now, clearly we get a different reading of (16), on this meaning of ‘because’ depending on whether that someone is Jane or John. The fact that we get the two readings of (16) that we do, one on which it entails (17)

and one on which it entails (18) suggests that the ‘someone’ in the meaning represented by ‘because_r’ can either be the ‘surface subject’, in this case Jane, of a passive sentence, which, in this case, is Jane, or the ‘underlying subject’ of the passive sentence, which, in this case, is John. But then the question arises as to why (15) does not also have the three readings. In order to see why, we need to see that what we have noted is a general phenomenon concerning adverbial expressions and phrases. Consider the following:

(19) John examined Jane reluctantly.

(20) Jane was examined by John reluctantly.

Now (19) has only one reading, but (20) has two, one on which it was John who was reluctant, and one on which it was Jane who was reluctant. Most linguists believe that this phenomenon arises because of the way the syntax of the adverbial phrase interacts with the syntax of the passive (Jackendoff, 1972; Thomason and Stalnaker, 1973; McConnell-Ginet, 1982; Wyner, 1998; Landman, 2000; Matsuoka, 2013). In short, passivisation makes it possible for the adverbial phrase to take either the ‘surface subject’ or the ‘underlying subject’ of the passive sentence as an argument. In contrast, in active sentences, the adverbial phrase must take the ‘surface subject’ as an argument, if it takes a subject as an argument at all.

In order to explain the two readings of (20), it is assumed that ‘reluctantly’ takes a subject as an argument, and that subject is the person who is reluctant about the relevant event. On this view, there is a tight interaction between syntax and semantics. While ‘reluctantly’ has a semantics on which it takes a subject as an argument, the syntax of the passive allows that argument to be either the ‘surface subject’ or the underlying subject. Importantly, it is *only* these subjects which can provide the needed argument.

To see this, consider an alternative hypothesis on which it is held that ‘reluctantly’ has an argument place which is filled by a contextually salient agent, rather than a syntactic subject. On such a view, we would expect to find a reading of either (19) or (20) on which some agent other than John or Jane was reluctant about John examining Jane or Jane being examined by John. Suppose, for example, that Jane’s mother doesn’t really want her to be examined by John, but she allows it to happen anyway. On this hypothesis we would expect a reading on which ‘Jane was examined by John reluctantly’ means that Jane was examined by John and

it was Jane's mother who was reluctant about it. But we do not find such a reading: 'Jane was examined by John reluctantly' is true only if either Jane was reluctant or John was reluctant.

Now, let's return to (16). If we apply the lessons of 'reluctantly' here, then we can hold that 'because_r' takes a subject as an argument. In (15) it must take the 'surface subject', which, in this case, is Jane, as the subject. In (16) it can take either the 'surface subject', which, in this case, is John, or the 'underlying subject', which, in this case, is Jane, as its argument. If it takes the former, then we get a reading which entails (18), and if it takes the latter, we get a reading which entails (17). So the fact that (16) has these two readings, is a good reason for thinking that 'because_r' has an argument place for a subject. This shouldn't be *too* surprising, given our initial, intuitive, gloss of the meaning of 'because_r'. In sentences like (1) it is obvious whose knowledge rationally explains John's going to the pub, namely, John's. But we can now see that it is a matter of the syntax and semantics interacting that this is so.

But this is not yet an argument for the conclusion that 'because' is lexically ambiguous. We want to argue that 'because' has a meaning, represented by 'because_r' on which it takes a subject as an argument, and that it also has a meaning, represented by 'because_c' on which it does *not* take a subject as an argument. If we could establish *this* then we will have established that 'because' is ambiguous, since, the difference between taking an argument and not is clearly a difference in meaning. Here we can simply point to the fact that a similar phenomenon does not arise in the case of passive sentences when the relevant reading is that which we would attribute to 'because_c.' Consider:

(21) The truck hit the tree because the road was wet.

(22) The tree was hit by the truck because the road was wet.

Absent any evidence for the claim that 'because' takes a subject as an argument in these contexts, we should conclude that it doesn't. We can conclude from this that 'because' is lexically ambiguous. But let's consider further evidence bearing on the issue.

3.8.2 Linguistic Tests for Ambiguity

There are several standard linguistic tests for the ambiguity of declarative sentences. These tests provide *prima facie* evidence for the presence or absence of ambiguity.

One linguistic test for the ambiguity of a declarative sentence is the contradiction test (Quine, 1960; Zwicky and Sadock, 1975, pp. 7–8; Gillon, 1990, pp. 407–410; Sennett, 2014, §4.3). On one version of the test, we ask whether a particular declarative sentence can be conjoined with its negation and nonetheless have a reading on which it is true relative to some context. If it can, then this suggests that it is ambiguous. If it cannot, then this suggests that it is not. Consider:

(23) John is going to the pub because he has a drinking problem, but he isn't going to the pub because he has a drinking problem.

(24) Jane wrote down the number 4 because $2 + 2 = 4$, but she didn't write it down because $2 + 2 = 4$.

If 'because' sentences were ambiguous between reason explanations and merely causal explanations, then we should expect there to be a reading of (23) on which it could be true, and a reading of (24) on which it could be true. My own view is that such readings are available. They can be brought out as follows:

(25) John is going to the pub because he has a drinking problem (since that's what people with drinking problems do), but he isn't going to the pub because he has a drinking problem (since he doesn't know that he has a drinking problem).

(26) Jane wrote down the number 4 because $2 + 2 = 4$, but she didn't write it down because $2 + 2 = 4$ (since the fact that $2 + 2 = 4$ doesn't causally explain anything).

Nonetheless, if it is possible to get these readings, it requires one to shift between the two different meanings of 'because'. This may not be so easy to do. We should distinguish between the prediction of the ambiguity hypothesis, which is that such sentences should have a reading on which it is possible for them to be true, and what we are asking for here. The hypothesis predicts that there *are* readings on which it is possible for such sentences to be true.

It doesn't predict that anyone should be able to interpret the sentence with the two meanings. It may be difficult to shift between the two meanings in interpreting the sentence.

Another kind of linguistic test for the ambiguity of a declarative sentence are so-called *identity tests*. Consider first, the *conjunction reduction test* (Sennett, 2014, §4.1). Suppose we have the two sentences:

(27) John is going to the pub because he has a drinking problem.

(28) Jane is going to the pub because John has a drinking problem.

We hypothesise that the 'because' in (27) is the purported 'because' of merely causal explanation and that the 'because' in (28) is the purported 'because' of reason explanation. Now, we consider the conjoined and reduced sentence:

(29) Both John and Jane are going to the pub because John has a drinking problem.

According to this test, (29) should be zeugmatic. It requires a 'crossed reading'. But a crossed reading is not permitted. This sentence strikes me as zeugmatic, which suggests our original sentences involve different readings, and are therefore ambiguous.

Consider a related test, the *ellipsis test* (Sennett, 2014, §4.2). Consider (27) and (28) again. Now consider the conjoined but elliptical sentence:

(30) John is going to the pub because he has a drinking problem, and Jane is too.

According to the test, (30) should be zeugmatic. This sentence strikes me as zeugmatic, which suggests our original sentences involve different readings, and are therefore ambiguous. These tests do better than the contradiction test, since they do not require us to interpret our sentences with different meanings. They only ask us to interpret it with one meaning.

I do not want to rest too much on the standard linguistic tests. That is why I have offered an independent argument for the ambiguity hypothesis. The main problem with the tests is that they require subtle judgements about whether particular sentences are zeugmatic. Such judgements are not particularly reliable. The different readings might be difficult to detect. At this point it is worth recalling Dray's remark that it is "...often possible to interpret an

explanation at the wrong level for a long time without committing any obvious logical errors” (Dray, 1960, p. 133). So it shouldn’t be surprising if the standard linguistic tests are inconclusive here.

3.8.3 *Multiplying Senses Beyond Necessity*

In his “Elements of Logic” (1855), Bishop Whateley writes:

... “Because” (*i.e.* “by-Cause”) is used to introduce either the Physical Cause or the Logical proof: and “Therefore,” “Hence,” “Since,” “Follow,” “Consequence,” and many other kindred words have a corresponding ambiguity: *e.g.* “the ground is wet, *because* it has rained;” or “it has rained, and hence the ground is wet;” this is the assignment of the *cause*: again, “it has rained, *because* the ground is wet;” “the ground is wet, and *therefore* it has rained:” this is assigning the logical *proof*; the wetness of the ground is the cause, not of the rain having fallen, but of *our knowing* that it has fallen (Whately, 1855, p. 373).

One often finds philosophers happily distinguishing between an ‘epistemic’ or ‘logical’ sense of ‘because’ and an ‘explanatory’ or ‘causal’ sense of ‘because’ (Morreal, 1979; Schnieder, 2010, 2011, p. 447). Strikingly, philosophers have been happier to accept an ambiguity here than with respect to non-psychological ‘because’-sentences. The argument for ambiguity here is much worse, since there are significant syntactic differences between the sentences. But, nonetheless, the question arises as to whether there is a third sense of ‘because’, then, distinct from those we have distinguished. Whately gives us a clue to the answer when he writes that the wetness of the ground is the cause of our knowing that the rain has fallen. Thus, we might think that the sense of ‘because’ involved in my utterance ‘It has rained, because the ground is wet’, is just the sense involved in your utterance of ‘He believes it has rained because the ground is wet.’ The contrast seems to be that I do not *assert* that I believe that it has rained because the ground is wet, but rather *express* my belief that the ground is wet, and *express* my reason for believing that it is wet. So accepting the conclusion of the argument of this section does not involve positing more senses of the word ‘because’ than are typically accepted.

3.9 The Context-Dependence Hypothesis

I suggested that an alternative explanation of the fact that someone can use (1) to say something which entails (2), and can also use it to say something which does not entail (2) is that (1) has a context dependent meaning, and means something which entails (2) in some contexts—we can call these psychological contexts—and means something which does not entail (2) in other contexts—we can call these merely causal contexts. Call this the *context dependence hypothesis*. According to this hypothesis (1) is not ambiguous, at least not in the way we have been supposing. In this section I want to ask whether ‘because’ could have a context dependent meaning which can explain why someone can use (1) to say something which entails (2), and can also use it to say something which does not entail (2). I will argue that it could not.

It is one thing to suggest that our facts might be explained in terms of context dependence, and quite another to give such an explanation. And it is not at all obvious how such an explanation would go. We can begin, perhaps, with the meaning of the ‘because’ of psychological explanation. What hypothesis can we offer about the meaning of the ‘because’ of psychological explanation such that we can explain how someone can use (1) to say something which entails (2), and can also use it to say something which does not entail (2)? Indeed, let’s begin with something even less ambitious. Let’s try to explain how (1) might be used to say something which entails that John knows that Jane is at the pub. Here is a promising strategy. We can assume that somehow or other, the ‘because’ of reason explanation places a contextually provided restriction on the explanantia of ‘because’ sentences. The thought then is that in psychological contexts, the restriction on the explanantia is to *reasons some contextually salient subject has for performing the action in question*. We can assume that the subject has such a reason only if it is something they *know* or *believe*. So, the sentence ‘John is going to the pub because Jane is there’ means, in such a context, roughly, that John is going to the pub because Jane is there *and* the fact that Jane is at the pub is a reason John has, and is therefore something he knows, for going to the pub, where ‘because’ here is the ‘because’ of psychological explanation or the ‘because’ of merely causal explanation. In merely causal contexts, the restriction on the explanans is relaxed. On this view, the meaning of ‘John is going to

the pub because Jane is there' is contextually dependent, it depends on some contextually salient restriction on the explanans. In this way we can explain how, relative to psychological contexts, (1) entails that John knows that Jane is at the pub, and in merely causal contexts, it does not.

But now we face a problem. It is not clear how to go from this hypothesis to one which explains how (1) could be used to say something which entails (2). The restriction on the explanantia would have to be something like: *a fact which is the content of a contextually salient subject's state of knowledge which explains why the subject is performing the action in question*. This is implausible. Moreover, it would lead to an implausible view of the meaning of (1), since, (1) would then have two explanantia: an agent's state of knowledge, and the content of that knowledge. But this is only the beginning of the problems for this hypothesis. Another problem is that it predicts more variation in the meaning of (1) than we in fact find. The only subject whose knowledge seems to matter is the subject of the main clause. Yet on the context dependence hypothesis, the knowledge of other contextually salient subjects should be able to be relevant, yet there are no contexts in which (1) entails that somebody other than John knows something (except for when the entailment is explained in other ways, for example, by involving an explanandum or explanans which entails that some other agent knows something). This suggests that the meaning of (1) is more tightly controlled by its syntax and semantics, in the way the ambiguity hypothesis predicts. Moreover, there is simply no independent linguistic evidence that (1) is context dependent in this way.

3.10 Conclusion

Do reason explanations entail psychological explanations? I have argued that they do. Are they a kind of psychological explanation? I have argued that they are.

I have argued that sentences like (1) have a reading on which they entail sentences like (2). More generally, I have argued that non-psychological 'because'-sentences have a reading on which they entail psychological 'because'-sentences. I have argued that the best explanation of this is that 'because' is ambiguous. Indeed, I have offered independent support for the claim. I offered the paradigm of the passive sensitivity of non-psychological 'because'-

sentences as good evidence that 'because' is ambiguous. This motivates the view that the meaning of 'because' involved in non-psychological 'because'-sentences which express reason explanations is roughly 'because x knows that', where 'x' is an argument place for a subject of the main clause, and 'because' here is the 'because' of psychological explanation. It follows, then, from the lexical semantics of 'because', on reasonable assumptions, that reason explanations are a kind of psychological explanation, and that psychologism about reason explanations is true.

CHAPTER 4

KNOWING WHY

4.1 Introduction

An important subproblem of the problem of self-knowledge is that of explaining how we know particular explanations of our own actions and attitudes. We seem to know such explanations in a non-inferential way, just as we seem to know particular facts about our own actions and attitudes in a non-inferential way. I seem to know, for instance, that I am going to the pub because Jane is there, in a non-inferential way. And I seem to know that I believe that Jane is at the pub because her car is in the car park, in a non-inferential way. If so, then the problem of explaining how we know particular explanations of our own actions and attitudes seems to be another subproblem of the problem of self-knowledge.

One good reason for thinking that we know such explanations in a non-inferential way is that we often know such explanations even though we do not have good reasons for believing that they obtain. As Donald Davidson writes:

Though you may, on rare occasions, accept public or private evidence as showing that you are wrong about your reasons, you usually have no evidence and make no observations. Then your knowledge of your own reasons for your actions is not generally inductive, for where there is induction, there is evidence.

(Davidson, 1980, p. 18)¹

The thought here is that in order to *know* particular explanations of our own actions and attitudes by inference our beliefs would have to be based on sufficient reasons for so believing. But since we do not seem to *have* sufficient reasons for so believing, our beliefs cannot be *based* on such reasons, and, therefore, we do not *know* particular explanations of our own actions and attitudes by inference.

If we know such explanations in a non-inferential way, then the question arises as to how we know such explanations, given that we know them in a non-inferential way, and the problem of explaining how looks like a subproblem of the problem of explaining how we know particular facts about our attitudes and actions, given that we know such facts in a non-inferential way.

Many theorists, however, deny that we know such explanations in a non-inferential way. They are impressed by particular self and other symmetries when it comes to knowledge of particular explanations of our actions. The symmetries, if they exist, are impressive. One particular symmetry concerns the similarity between the explanations we give of our own actions and attitudes and the explanations given by suitably placed observers who must make inferences about the explanations of our actions and attitudes on the basis of evidence that they have. It turns out that we are prone to make the same mistakes as the observers. This is a good reason for thinking that we do not know such explanations in a way which is radically different to how others know such explanations, that is, in a non-inferential way. And it is a good reason, it seems, for thinking that we know such explanations by inference. The thought here is that the best explanation of the observed self and other symmetries is that both must be arriving at their explanations by inference.²

¹James Pryor writes "...your beliefs about the reasons for which you acted are not ordinarily based on any evidence" (Pryor, 2005, p. 533). For similar claims see: (Setiya, 2013, p. 192; Sandis, 2015; Baker, 2015, p. 3046).

²This argument is found in the work of Richard E. Nisbett and Timothy DeCamp Wilson (Nisbett and Wilson, 1977). Timothy Wilson, summarising the work of Nisbett and himself many years later writes that people "...do not have privileged access to the causes [of their responses] and must infer them" (Wilson, 2002, p. 106). See also (Nisbett and Ross, 1980, p. 223). More recently, Shaun Nichols and Stephen Stich have written: "...on our account, when trying to figure out the *causes* of one's own behaviour, one must reason about mental states, and this process is mediated by [a body of information about the mind]" (Nichols and Stich, 2003, p. 163). Discussing various empirical studies, Eric Schwitzgebel writes that the psychological literature "...can

Many theorists have been persuaded by this simple argument and are happy to narrow the scope of the problem of self-knowledge to that of explaining how we know particular non-explanatory facts about our own actions and attitudes.

So, we seem to have good reasons for thinking that we know particular explanations of our actions and attitudes in a non-inferential way *and* good reasons for thinking that we do not. We have a problem. How might we respond? Perhaps we *do* have good reasons for believing that particular explanation of our actions and attitudes obtain after all: perhaps Davidson is wrong. Alternatively, perhaps on closer examination we find that the self and other symmetry evaporates: perhaps the empirical evidence is misleading. Or perhaps we can make the claim that we know particular explanations of our actions and attitudes in a non-inferential way consistent with the observed self and other symmetry. That is, perhaps the self and other symmetries do not support the inferential hypothesis, as we might call it, over some alternative non-inferential hypothesis.

In this essay I defend this third option. I begin by laying out Nisbett and Wilson's argument for the conclusion that we must know particular explanations of our own actions and attitudes by inference and then examine the experimental evidence their argument draws on. I argue that, although the evidence is not entirely unproblematic, we should tentatively accept it, and accept that there are certain symmetries between self and other when it comes to explanations of our own actions and attitudes. I then examine the reasons we have for thinking that we know particular explanations of our own actions and attitudes in a non-inferential way. At this point we are left with a serious tension: we have what looks like a good argument for a conclusion that we simply cannot accept, since we also have good reasons for not believing that conclusion. What we need is an alternative explanation of Nisbett and Wilson's symmetry and the features of our knowledge of explanations of our actions and attitudes. In the final section I develop such an alternative according to which our knowledge of particular explanations of our actions and attitudes is arrived at by resolving particular practical questions about our reasons for our actions and attitudes. I argue that this theory can explain the self and other symmetry and the other features of our knowledge of explanations of our

be interpreted as suggesting that the causes of our behaviour are not, after all, the sorts of things to which we have introspective access" (Schwitzgebel, 2014, §4.2.1).

actions and attitudes. An appendix discusses further empirical evidence which bears on the question. I tentatively argue that the evidence supports our alternative explanation.

Before turning to the argument, a few further remarks. First of all, I should note that I would be happy if it turned out that the empirical evidence did not support the self and other symmetry. That would undercut the main argument for thinking that our knowledge of particular explanations of our own actions and attitudes is inferential. I have tried to evaluate that evidence in this essay. When I suggest that we tentatively accept that the evidence supports the self and other symmetry, I really do mean that we should *tentatively* accept it. But to the extent that we are only tentatively accepting this conclusion, I can hardly use the conclusion in an argument *for* the alternative theory of how we know such explanations. My thought here is that I can cover both bases: either the symmetry is not supported, in which case the argument for the inferential hypothesis is undercut, or the symmetry is supported, but is better explained by an alternative non-inferential hypothesis. Either way, I think there is no good argument for the inferential hypothesis. But insofar as I do not take a firm stand on whether the symmetry is supported I do not have an argument *for* the alternative explanation I offer. For the moment I am happy to live with that consequence. This is not the only argument that could be offered for the alternative explanation I want to offer. Perhaps the best independent argument for the alternative account I offer is that there simply are no other (remotely plausible) alternative explanations of how we know particular explanations of our own actions and attitudes in a non-inferential way.

4.2 The Argument from Symmetry

In this section I examine Nisbett and Wilson's influential argument for the conclusion that we know particular explanations of our own attitudes and actions by inference (Nisbett and Wilson, 1977). Strikingly, although Nisbett and Wilson's essay is often cited in support of this conclusion, the particular arguments of the essay are rarely discussed in any detail. In what follows I will carefully examine both the structure of the argument and the support given for the crucial premises.

The starting point for Nisbett and Wilson's argument is a particular kind of self and other

symmetry. The particular kind of self and other symmetry which interested Nisbett and Wilson was a symmetry in kinds of error and ignorance about particular explanations of actions and attitudes between self and other. Nisbett and Wilson claimed on the basis of their experimental studies that the kind of errors and the kind of ignorance we make and have about particular explanations of our *own* actions and attitudes are just the same kind of errors and ignorance someone else would make and have about particular explanations of our actions and attitudes. They argued that the best explanation of this symmetry is that we must be arriving at explanations of our own actions and attitudes by inference, just like someone else must. They assumed, plausibly, that others must arrive at explanations of our actions and attitudes by inference. So a good explanation of why we make the same kind of errors and have the same kind of ignorance about explanations of our own actions and attitudes is that we are arriving at our explanations by inference from roughly the same evidence.

In developing this argument Nisbett and Wilson are drawing on a similar argument which had been made by Daryl Bem concerning knowledge of our attitudes (Nisbett and Wilson, 1977, p. 248; Bem, 1967). Bem had argued that we showed the same kind of errors and ignorance about our own attitudes as a suitably placed observer, and inferred from this that we must be arriving at beliefs about our own attitudes by inference from roughly the same evidence as observers. Bem's argument, in the case of ascriptions of certain attitudes is roughly the following: the self-ascriptions of an agent of attitudes to herself and the ascriptions of an external observer with access to much the same information as the agent of attitudes to the agent are relevantly similar; the best explanation of this is that the agent does not have a "fount of privileged self-knowledge", rather, the agent's self-ascriptions of attitudes "may be viewed as inferences from observations of his own overt behavior and its accompanying stimulus variables"; so agents are arriving at self-ascriptions by inference (Bem, 1967, p. 186). Bem goes on to speak of the "... empirical generalization that an individual's belief and attitude statements and the beliefs and attitudes that an outside observer would attribute to him are often functionally similar in that both sets of statements are partial "inferences" from the same evidence ..." (Bem, 1967, p. 186). Here are Nisbett and Wilson making just this argument in the text:

... whatever capacity for introspection exists, it does not produce accurate reports about stimulus effects, nor does it even produce reports that differ from predictions of observers operating only with a verbal description of the stimulus situation. . . . [I]f the reports of subjects do not differ from the reports of observers, then it is unnecessary to assume that the former are drawing on “a font of privileged knowledge.” It seems equally clear that subjects and observers are drawing on a similar source for their verbal reports about stimulus effects. (Nisbett and Wilson, 1977, pp. 247–248), quoting from (Bem, 1967, p. 186)

Introspection

Since the argument is an inference to the best explanation, it is important to be clear about what the alternative explanations are. Nisbett and Wilson assume that the only relevant alternative explanation is one on which we know particular explanations of our own actions and attitudes by *introspection*—on an observational model of introspection. Nisbett and Wilson do not offer an explicit technical definition of introspection, but they do give us some hints about what they had in mind. They speak of “direct introspective access” and “direct introspective awareness” and speak of the potential ability of subjects to “... observe directly the workings of their own minds” (Nisbett and Wilson, 1977, p. 232).³ The latter suggests that they have something like an observational or perceptual theory of introspection in mind which involves a kind of inner observation or perception. They seem to have had in mind a picture on which our knowledge of particular explanations of our own actions and attitudes involved a kind of inner awareness of the particular mental processes relevant to such explanations. This is suggested by the following question they ask early in the essay: “If it is not direct introspective access to a memory of the process involved, what is the source of such verbal reports?” (Nisbett and Wilson, 1977, p. 232). And it is suggested again later in the essay when they answer this question:

³As White points out the literature on Nisbett and Wilson is a terminological quagmire. He points out that Nisbett and Wilson themselves use: ‘introspective access’, ‘aware/unaware’, ‘direct access’, observe [the workings of their own minds]’, ‘conscious awareness’, ‘interrogate [a memory of]’, ‘introspective awareness’, ‘conscious’, ‘[hidden from] conscious view’, ‘genuinely insightful introspection’, ‘knowledge of’, and ‘consulting [a memory of]’ (White, 1988, p. 17).

We propose that when people are asked to report how a particular stimulus influenced a particular response, they do so not by consulting a memory of the mediating process, but by applying or generating causal theories about the effects of that type of stimulus on that type of response. (Nisbett and Wilson, 1977, p. 248)

Thus, Nisbett and Wilson seem to have an observational model of introspection in mind, one on which we are directly introspectively aware of the mental processes relevant to explanations of our actions and attitudes. (See (Lyons, 1986) for a discussion of history of this mode of introspection. And see (Shoemaker, 1996) for some critical discussion). There are many ways one can develop such an account of introspection. For now we can simply assume that the process is some kind of perception-like process by which one becomes aware of one's mental processes, where this involves, at a minimum, some kind of direct causal process by which one becomes aware of one's own mental processes (Goldman, 2006, p. 246). This is the main alternative hypothesis they consider.

It may seem striking that Nisbett and Wilson consider a kind of inner-sense model of introspection as an alternative model of how we know particular explanations of our own actions and attitudes. According to Alvin Goldman:

No careful privileged-access theorist should claim that people have introspective access to the causes of their behavior, especially qua causes, or to the cognitive processes that run their mental lives. (Goldman, 2006, p. 233)

This awareness-of-causes thesis, however, is one that no classical introspectionist, to my knowledge, has ever asserted. (Goldman, 1993, p. 27)

Goldman and others think that the implausibility of an inner-sense model of introspection can be established *a priori*. He writes:

As we have known since Hume (1748), causal connections between events cannot be directly observed; nor can they be introspected. A sensible form of introspectionism, therefore, would not claim that people have introspective access to

causal connections, but this leaves it open that they do have introspective access to the mere occurrence of certain types of mental events. (Goldman, 1993, p. 27)

If the inner-sense model of introspection holds that we know particular explanatory facts about our own actions and attitudes by a kind of inner-perception, and it is impossible to know such facts by inner-perception, then the model is implausible. But it isn't at all obvious that we cannot know causal facts by sense perception (Beebe, 2003; Siegel, 2010) and it isn't at all obvious that the kind of particular explanatory facts about our own actions and attitudes are *ordinary* causal facts.

Moreover, Nisbett and Wilson do not say that the inner-sense model of introspection is committed to the claim that we can know particular explanatory facts about our own actions and attitudes by perception. Rather, they seem to think that what we know by a kind of inner-perception is that some higher-order cognitive process has taken place. If there is a criticism of Nisbett and Wilson lurking here it is that they do not say how they think that knowledge of such cognitive processes connects with knowledge of particular explanations of our actions and attitudes. But it is hard to see that this amounts to much of a criticism. If one is perceptually aware, somehow, that one is *inferring p from q* then it is only a small gap to knowledge that one believes that p because q.

Inference

Nisbett and Wilson similarly do not offer a technical definition of inference, so we can assume that they intend their conclusion to be about the ordinary notion of inference. According to that notion, inferring is a kind of mental process, and in the first instance, one infers one fact or facts from another fact or set of facts. I might, for instance, infer that John is home from the fact that his lights are on. I infer one fact, namely, the fact that John is home, from another fact, namely, the fact that his lights are on. But if I infer that John is home from the fact that his lights are on, I can also be said to infer that John is home from my *belief* that he is home. In general, it seems that one can infer some fact from another fact only if one infers the former fact from a belief whose content corresponds to the latter fact. The converse doesn't hold, however. I might infer that John is home from my belief that his lights are on without,

thereby, having inferred that John is home from the fact that his lights are on, since his lights might not be on. The important point for now is that according to the ordinary notion of inference, one can infer something from some fact only if one infers it from a belief whose content corresponds to that fact. When we do, we can say that one belief is psychologically based on the other.

It is also part of our ordinary notion of inference that in order to *know* something by inference, one must infer that thing from *sufficient* reasons for believing it. In order for me to know that John is home by inference I must infer that he is home from sufficient reasons for believing it. If I infer that he is home from the fact that his lights are on, among other facts, then I may well count as having inferred that he is home from sufficient reasons. These truisms about inference will be relevant later, for one can, of course, respond to some of the difficulties I raise for Nisbett and Wilson by arguing that we know explanations of our actions and attitudes by inference on a difference conception of inference, one which departs either psychologically, or epistemologically, or in both ways, from the ordinary notion.

Summing Up

Understood this way, it is clear that the inferential hypothesis, as we might call it, provides a better explanation of symmetries between self and other, than the introspection hypothesis. The introspection hypothesis provides no explanation at all of the symmetries between self and other. Why should we expect such symmetries if we know particular explanations of our own actions and attitudes by introspection thus conceived? We should expect asymmetries.

Indeed, it is difficult to see why we should expect such symmetries between self and other on any hypothesis other than the inferential hypothesis. The process by which we arrive at explanations of our own actions and attitudes *must* be sensitive to the same kind of things that the explanations of our own actions and attitudes of a suitably placed observer if the symmetry in the pattern of errors and ignorance between self and other is not to be a miracle. So it doesn't matter too much that Nisbett and Wilson were vague about alternative hypotheses. Any hypothesis in the vicinity of the perceptual or observational model of introspection is going to fail to offer a process which is sensitive to the same kind of things as the observer's

process.

It will be helpful at this point to lay out Nisbett and Wilson's argument explicitly. We can think of their argument as having roughly the following form:

The Major Empirical Premise: Typically, the beliefs of actor subjects about rational explanations of the actions and mental states of actor subjects are false if, and only if, the beliefs of observer subjects about those explanations are false. (Supported by experiments)

The Minor Empirical Premise: Observer subjects arrive at beliefs about rational explanations of the actions and mental states of actor subjects by inference from other things they know. (Obvious)

Conclusion: Therefore, actor subjects arrive at beliefs about rational explanations of their own actions and mental states by inference from other things they know. (Inference to the best explanation from the Major Empirical Premise and the Minor Empirical Premise.).

What follows will be structured around this argument. For the rest of this section I will evaluate the evidence for the major empirical premise. I take the minor empirical premise to be uncontroversial. I will primarily be concerned with the question of how good the inference is. As we have just seen, if inference and inner sense are the only relevant alternative explanations, then we seem to have very good reasons for thinking that our knowledge of particular explanations of our own actions and attitudes is inferential. These reasons may be better than any reasons we have against believing this. I am going to argue, however, that the plausibility of the argument is undermined once we see that there is a viable alternative non-inferential explanation available.

I have attributed to Nisbett and Wilson a kind of inference to the best explanation. Strikingly, this is *not* the argument often attributed to Nisbett and Wilson by theorists. Often, theorists attribute to Nisbett and Wilson a kind of argument from the falsity of infallibility of omniscience to the falsity of the introspection hypothesis and to the truth of the inferential hypothesis. The thought it is that Nisbett and Wilson's experiments demonstrate that we are

neither infallible or omniscient about the explanations of our own actions and attitudes, so we do not know them by introspection on the inner-sense model, and must know them by inference. Now, this is a manifestly bad argument, and it is not Nisbett and Wilson's. First of all, the argument assumes that there is a connection between the inner-sense model of introspection and infallibility and omniscience. But it has long been acknowledged that there is no such connection (Lyons, 1986). One can defend an inner-sense model of introspection without being committed to infallibility and omniscience. Secondly, the argument assumes that the only alternative to the inner-sense model is an inferential model. This assumption is also questionable.

But this brings us to an argument which might more reasonably be attributed to Nisbett and Wilson. And this is an argument by elimination. It might be thought that Nisbett and Wilson are arguing that the only alternatives here are inference and introspection on the inner-sense model. It might be thought that their argument against the inner-sense model of introspection is that our explanations of our own actions and attitudes are too unreliable to be explained on the model of introspection, so, therefore, the inferential hypothesis must be correct. This is closer to Nisbett and Wilson's actual argument, but, as we have seen, it is clearly *not* their argument. It is true that Nisbett and Wilson seem to think of the inner-sense model and the inference model as the only alternatives. But they do so within the context of an argument to the best explanation, not an argument by elimination. It is plausible to assume that the inner-sense model of introspection entails some kind of epistemic advantage which Nisbett and Wilson's experiments might bear on. And Nisbett and Wilson often write in ways which suggest that they think that they show that our explanations of our own actions and attitudes are too unreliable to be explained on the basis of the inner-sense model. This is plausible. But it doesn't follow that Nisbett and Wilson have an argument by elimination in mind. Rather, they are just pointing to a reason against thinking that the inner-sense model is correct. The overall structure of their argument remains an inference to the best explanation.

4.2.1 *The Evidence for Symmetry*

Nisbett and Wilson attempt to establish experimentally that there are particular *symmetries* between self and other when it comes to explanations of actions and attitudes. Those symmetries involve particular patterns of error and ignorance about the explanation of an agent's actions and attitudes. Summarising their findings, they write:

...the evidence suggests that people's erroneous reports about their cognitive processes are not capricious or haphazard, but instead are regular and systematic. Evidence for this comes from the fact that "observer" subjects, who did not participate in the experiments but who simply read verbal descriptions of them, made predictions about the stimuli which were remarkably similar to the reports about the stimuli by subjects who had actually been exposed to them. (Nisbett and Wilson, 1977, p. 247)

The experimental setup by which Nisbett and Wilson sought to establish this was simple. Nisbett and Wilson had 'actor-subjects' make particular choices, decisions, and evaluations, and then asked those subjects why they made the choices, decisions, and evaluations, that they did. Nisbett and Wilson had so-called 'observer-subjects', who were given information about the actor subject, their situation, and their choice, explain the actor subject's choice, decision, or evaluation.

What Nisbett and Wilson claim to have found is no significant difference in the explanations given by actor subjects and observer subjects regarding the actor subject's choice, decision, or evaluation. When actor subjects were in error about a particular explanation, so too were the observer subjects, and vice versa. When actor subjects were ignorant of a particular explanation, so too were the observer subjects, and vice versa. In short "...subject reports were accurate if and only if observer predictions were also accurate" (Nisbett and Bellows, 1977, p. 622). From this Nisbett and Wilson inferred that actor subjects must be arriving at explanations of their own actions and attitudes by inference. Here is how Nisbett and Nancy Bellows put it: "The similarity of subject reports about the effects of the factors and observer predictions about their effects was so great as to indicate that they must have

been generated by similar processes, that is, by reliance on similar a priori causal theories” (Nisbett and Bellows, 1977, p. 622).

Let’s briefly review several of the studies from which Nisbett and Wilson draw this conclusion. In one study, which we can call the emotional impact of literature study, subjects were presented with a passage of literature in which an innocent child is drowned and then asked about the emotional impact particular parts of the passage had had on them (Nisbett and Wilson, 1977, p. 245).⁴ Nisbett and Wilson had found that the absence or presence of particular parts of the passage typically had no effect on the emotional impact of the passage or the reported emotional impact of the passage. However, subjects reported that the relevant parts of the passage increased the emotional impact of the passage. So actor subjects seemed to be in error about the explanation of their responses.

Nisbett and Wilson also presented the relevant passages to subjects who did not originally receive them and asked these subjects to predict what the emotional impact of these passages *would have been*, had they been presented with them originally. In other words, Nisbett and Wilson had observer subjects give explanations of the actor subjects’ emotional response. Strikingly, the answers given by the observer subjects were basically the same as those of the subjects who were originally presented with the passages (Wilson and Nisbett, 1978, p. 126). The observer subjects seem to make a similar error, since it is plausible to assume that the observer subjects would not themselves have had such responses themselves.

In another study, which we can call the reassurance and willingness study, subjects were reassured that the electric shocks they were receiving were not doing them permanent damage, and then given increasingly higher shocks (Nisbett and Wilson, 1977, p. 246).⁵ Nisbett and Wilson claim that whether subjects receive reassurances or not does not make a difference to how far the subjects went in taking shocks. Subjects were then asked whether the fact that they received the reassurance explained why they had gone as far as they had. Subjects typically believed that it did. So subjects seem to be in error about an explanation of their action.

⁴Discussed briefly in (Nisbett and Ross, 1980, p. 209) and at greater length in (Wilson and Nisbett, 1978, pp. 124–126).

⁵Briefly discussed in (Nisbett and Ross, 1980, p. 210) and at greater length in (Wilson and Nisbett, 1978, pp. 126–128).

Nisbett and Wilson report that observer subjects who were merely asked to predict whether the reassurance would make a difference to how much shock a subject would take and whether it would increase their shock, made basically the same predictions: "...control subjects' predictions about the effects of inclusion of the phrase were similar to the incorrect reports of experimental subjects. Half believed the phrase would have had an effect and of these, two thirds believed the phrase would have increased their willingness to take shocks" (Wilson and Nisbett, 1978, p. 128). The observer subjects too seemed to be in error. These subjects would offer explanations of the actions of actor subjects which did not in fact obtain.

In the above studies the inclusion of an observer subject was something of an afterthought. In another study, however, Nisbett and Bellows set out explicitly to establish the symmetry between self and other concerning explanations of attitudes and actions (Nisbett and Wilson, 1977, pp. 249–250; Nisbett and Bellows, 1977).⁶

Nisbett and Bellows' study used female college students and concerned the reasons for which they made certain judgements. The subjects were given information about a similarly aged, female, job candidate and were asked to make certain judgements about the candidate. Different groups of subjects were given slightly different information about the candidate. Half were told the candidate was attractive and the other half were not, a different half were told that she had an excellent academic record and the other half were not, a different half again were told that the candidate had spilt coffee at the interview and the other half were not, a different half again were told that the candidate had been involved in an auto accident and the other half were not, and a different half again were told that they would have an opportunity to meet the candidate and the other half were not (Nisbett and Bellows, 1977, p. 615). The judgements the subjects were asked to make were in answer to the following questions: (1) How much do you think you would like this person?; (2) How sympathetic would this person be to the feelings of others?; (3) How intelligent do you think this person is?; (4) How flexible would the person be in solving problems? The subjects were then asked to answer questions concerning "how certain factors influenced their judgements" (Nisbett and Bellows, 1977, 618). They were then asked questions like: "How did the person's aca-

⁶For further discussion see (Nisbett and Ross, 1980, pp. 212–216). See (Smith and Miller, 1978, p. 358) for a critical discussion of some of the inferences involved in this study.

ademic credentials influence your judgement of how intelligent the person is?” Such questions can be seen as asking whether facts about the person’s academic credentials, say, were among the reasons why the agent made the judgement she did about the intelligence of the person. Another group of female students, the so-called “observer subjects” were “told that the investigators were interested in how people make judgements about others from particular kinds of information” (Nisbett and Bellows, 1977, p. 618). These subjects were not given any general information about the candidate, but were told to imagine that the candidate is of the same age and sex as themselves. Observers were then asked to “suppose you knew the person was quite physically attractive”, for example, and then asked questions like “How would that influence how much you like the person?” (Nisbett and Bellows, 1977, p. 618). Nisbett and Bellows predicted that (i) subject reports would be accurate concerning the influence of information about academic performance on their judgements of intelligence, but inaccurate with respect to the influence of information about meeting the candidate on their judgements of likability; (ii) the accuracy of subject reports would not exceed the accuracy of observer predictions; (iii) and observer predictions would be highly similar to actor reports. Here is Nisbett and Bellows’ report of what they found:

Subject accuracy did not exceed observer accuracy for any of the judgements. For the more subjective judgements, the accuracy of subject reports was nil and so was the accuracy of observer predictions. For the intelligence judgement, subject accuracy was virtually perfect. This accuracy cannot be attributed to introspective capacity on the part of the subjects, however, because the accuracy of observer predictions equalled the accuracy of subject reports. The correlation between observer predictions and actual effects (.98) was trivially higher than the correlation between subject reports and actual effects (.94). Thus subject reports were accurate if and only if observer predictions were also accurate. (Nisbett and Bellows, 1977, pp. 621–622)

This evidence certainly seems to support the self and other symmetry. If these experimental results generalise, then it certainly looks like “[subjects’] explanations about the causes of their response are no more accurate than the explanations of a complete stranger who lives in

the same culture” (Wilson, 2002, pp. 108–109).

4.2.2 Problems With the Evidence for Symmetry

It is on the basis of studies like these that Nisbett and Wilson take themselves to have established that there is a particular kind of *symmetry* between self and other when it comes to explanations of actions and attitudes. There are many questions that we can and should ask about these studies. Indeed, many such questions have been asked.⁷ Two particular concerns stand out.

One issue is the following. So far the symmetry claim has been presented as the claim that for any particular explanation which obtains or does not obtain, individual actor subjects are just as likely to be in error about that particular explanation as individual observer subjects. Of course, in order to directly test a claim like this, one would need a way of knowing when an explanation obtains and when one doesn’t, and one would need a way of knowing what the subject believes about explanations. It is notoriously difficult to know whether an explanation obtains in any particular case.

It turns out that Nisbett and Wilson did not directly test this claim, however. Rather, they tested the claim that for any particular explanation of the average response of the actor subjects, taken together, the average response of actor subjects, taken together, is just as likely to be erroneous, as the average response of observer subjects, taken together. For instance, in the reassurance and willingness study, receiving a reassurance about the shock did not have an influence on the *average* willingness to take shocks. But, on average, both actor subjects and observer subjects believed that the actor subjects went as far as they did in taking shocks because they received the relevant reassurance. (Likewise, in the job applicant study, what Nisbett and Bellows found was that groups of actor subjects were correct about the influence of some factor (on the group of actor subjects) just in case groups of observer subjects were. This is made clearer in (Nisbett and Ross, 1980)).

The obvious problem with this kind of between-subject design is it that is compatible

⁷The critical literature on Nisbett and Wilson’s major premise is extensive. See (Ericsson and Simon, 1980; Howe, 1991; White, 1988; Kraut and Lewis, 1982; Sabini and Silver, 1981; Smith and Miller, 1978; Sprangers et al., 1987; Wright and Rip, 1981; White, 1980, 1987) and especially (White, 1988).

with the reassurance having no influence on the actor subjects, that, say, half of the actor subjects *were* influenced by the reassurance, and half were not. And it is compatible with the average of the actor subjects' reports being that there was an influence, with, say, the half of the subjects who *were* influenced by the reassurance, reporting that they were, and the half of the subject who were not influenced by the reassurance, reporting that they were not.

As Nisbett and Ross write:

The between-subject design of the study make it impossible to assess the accuracy of the individual subjects' causal reports. It could be that, despite the failure of subjects as a group to distinguish between effective and ineffective manipulations, particular subjects may have reported accurately the influences on their own judgements. (Nisbett and Ross, 1980, p. 215)

Several critics of Nisbett and Wilson have drawn attention to this (Smith and Miller, 1978; White, 1980, pp. 109–110; Wright and Rip, 1981, p. 602).⁸ But the mere compatibility between the average response and answer with the lack of error on the actor subject's part, does not show much. This is just an alternative hypothesis. Surely the *best* explanation of the symmetry that Nisbett and Wilson did directly test is that for any particular explanation which obtains or does not obtain, individual actor subjects are just as likely to be in error about that particular explanation as individual observer subjects.

Another concern about Nisbett and Wilson's studies is more pressing. This concern is that Nisbett and Wilson do not carefully distinguish between different *kinds* of explanations of our actions and attitudes (Smith and Miller, 1978, p. 357; White, 1988, p. 21). Nobody

⁸Kraut and Lewis write: "The most important criticism is that the data from this and other studies conducted by Nisbett and his colleagues, which used between-subject designs, are irrelevant to questions of self-awareness. The self-awareness question is one about an individual: How accurate is he or she at assessing the influences on his or her beliefs, decisions, or behavior? This question cannot be answered by showing that a group member identifies or fails to identify the factors that influence the group" (Kraut and Lewis, 1982, p. 449). Moreover, (Smith and Miller, 1978) offer a reanalysis of the data from (Nisbett and Bellows, 1977). They argue that on this reanalysis: "there is a substantial and certainly significant evidence for introspective self-awareness on the part of subjects in Nisbett and Bellow's own study: Those subjects whose rating actually was above the mean and vice versa" (Smith and Miller, 1978, p. 358). They claim that similar reanalyses of the data from other experiments have the same results: "A similar criticism of the data analysis applies to several other studies: those concerning the effect of distraction on ratings of movies, the emotional impact of literary passages, and the effects of reassurance on willingness to take electric shock" (Smith and Miller, 1978, p. 359). See (Nisbett and Ross, 1980, p. 215) for a response. See also (Howe, 1991).

thinks that we are able to know *all* kinds of explanations of our actions and attitudes in a non-inferential way. Our actions and attitudes have ever so many different kinds of explanations. They have historical, neurological, and micro-physical, explanations, but we cannot know *these* explanations non-inferentially, and nobody claims that we can. Rather, what is at issue is whether we can know what we might call *rational explanations* of our actions and attitudes, which include what we might call *reason explanations*—explanations whose explanantia are the agent’s reasons for her action or attitude—and *psychological explanations*—explanations whose explanantia are psychological states which rationalize the agent’s actions or attitude (Alvarez, 2010).

Now, it may seem that Nisbett and Wilson are insensitive to these distinctions. To some extent that is so. Nisbett and Wilson seem to be interested in both *merely psychological explanations*, which aren’t necessarily a kind of rational explanation, and *psychological explanations proper*, which are a kind of rational explanation.⁹ Recall the job applicant study. Nisbett and Wilson found that a subject’s belief that she was about to meet a particular job applicant explained her evaluation of the applicant’s likeability. Subjects denied that this was so. So they seemed to be in error. But, arguably, they are not in error if what they are denying is that a particular *rational* explanation obtains with this belief as its explanans. They may well be right about this although a merely psychological explanation may obtain.

In light of this, one might argue that once we exclude ignorance and error of merely psychological explanations and merely causal explanations from the data, we will no longer find a symmetry between self and other. An objection of this kind has been put forcefully by Constantine Sandis. He writes:

A worry with [Nisbett and Wilson’s] analysis is its lack of any distinction between the causes of bodily behavior and agential reasons for acting. Ironically, their argument unintentionally suggests that laypeople might be making just such a distinction. If so they would be right to do so: the position of a pair of stockings on a table is rarely, if ever, a reason for which one chooses them over another

⁹This makes sense of their interest in the study by (Nisbett and Schachter, 1966), reported in (Nisbett and Wilson, 1977, p. 327), and in the study by (Storms and Nisbett, 1970), reported in (Nisbett and Wilson, 1977, pp. 237–238). This also makes sense of Nisbett and Wilson’s interest in subliminal perception and problem solving (Nisbett and Wilson, 1977, pp. 239–241).

pair. It could, however, explain why we mistakenly come to think of them as being smoother etc. What we are fabricating in such a case is not a tale about our agential reasons but one about the quality of the stockings (Sandis, 2015, p. 270).

I have some sympathy with this objection. But Sandis has cherry-picked his case. True, Nisbett and Wilson miscategorise this case. But we can reexamine the studies with this distinction in mind and see whether their experiments do show the relevant pattern of errors in the case of *rational* explanation. If we look carefully at Nisbett and Wilson's studies, we see that in many cases the explanations involved *are* rational explanations, and that there is symmetry between self and other with respect to such explanations. For example, it is very plausible that in the job applicant study, particular rational explanations obtain which the subjects deny obtain and that the explanations the subjects claim obtain do not obtain.

None of this is to say that Nisbett and Wilson's studies are unproblematic. I think it would be a mistake, however, to dismiss them out of hand, and optimistically hope that future research will not establish a symmetry between self and other with respect to such explanations. Indeed, there is evidence from split-brain studies, and choice-blindness studies, which suggest that subjects will explain choices that they did not make, in something like the way the subjects in these experiments offer explanations for responses which have alternative psychological explanations (Gazzaniga and LeDoux, 1978; Gazzaniga, 1995; Johansson et al., 2005, 2006). (I discuss these studies in the appendix). I suggest that we do not leave our evaluation of Nisbett and Wilson's argument open to empirical hostage in this way. So I propose to tentatively accept that Nisbett and Wilson have established that there is a kind of symmetry between self and other when it comes to rational explanations of our actions and attitudes. My strategy will be to argue that even if we grant Nisbett and Wilson this assumption, we should not accept their conclusion. That is, we should not accept that our knowledge of particular explanations of our own actions and attitudes is inferential.

4.3 Against Inference

We have good reasons for thinking that our knowledge of particular explanations of our own actions and attitudes is non-inferential. So, we cannot simply accept Nisbett and Wilson's

conclusion. We face a problem which needs to be resolved. Strikingly, Nisbett and Wilson seem to think that the only reasons *against* the hypothesis that our knowledge of such explanations is inferential is that it does not *feel* to us like we are making an inference when we arrive at such explanations. They attempt to provide an alternative explanation of this fact. But this isn't the only reason for thinking that our knowledge of such explanations is non-inferential. As we saw above, the fact that we often simply do not have adequate reasons for believing that such explanations obtain is a good reason for thinking that we do not know such explanations by inference. This was Davidson's point. But there are further reasons for thinking that we do not arrive at such knowledge by inference. In this section I isolate some of these reasons.

It is helpful at this point to remind ourselves of the features of our ordinary notion of inference which we isolated above. The two central features were that if we infer something from some fact we infer it from a belief whose content corresponds to that fact and that in order to know something by inference we must infer it from sufficient reasons for believing it. There are good reasons for thinking that our knowledge of particular explanations of our own actions and attitudes is not arrived at by inference in this sense. We don't infer it from other things we believe, and we simply do not seem to have sufficient reasons for believing such explanations. Of course, one might reject the ordinary notion of inference. But that is to change the subject. The question is whether such knowledge of particular explanations of our actions and attitudes is non-inferential on the ordinary notion of inference.

If we arrive at knowledge of particular explanations of our actions and attitudes, then we must infer those explanations from sufficient reasons for believing that the explanations obtain. But our explanations do not seem to be based on sufficient reasons, because we do not seem to have sufficient reasons for believing such explanations. This is Davidson's point when he writes: "your knowledge of your own reasons for your actions is not generally inductive, for where there is induction, there is evidence" (Davidson, 1980, p. 18). Suppose I am waiting in my office because I am expecting a phone call. I might know that I am waiting in my office and that I am expecting a phone call. But these facts, taken together, do not initially seem to be sufficient reasons for believing that I am waiting in my office because I am

expecting a phone call. They might be good enough reasons for *suspecting* that this is why I am waiting in my office, but not for *believing* it. So I can't *know* that I am waiting in my office because I am expecting a phone call on the basis of such reasons.

But not so fast. Can't some suitably placed observer come to know that I am waiting in my office because I am expecting a phone call on the basis of the fact that I am waiting in my office and I am expecting a phone call? Would it be a reasonable inference to the best explanation that that is why I am waiting, especially in the absence of alternative explanations of the fact that I am waiting in my office. Indeed, the point must be conceded. I *do* seem to have sufficient reasons for believing that I am waiting in my office because I am expecting a phone call.

But we have described an atypical case, one in which I *already* know that I am waiting in my office and that I am expecting a phone call. Typically, the fact that I am waiting in my office will not be a premise from which I infer that I am waiting in my office because I am expecting a phone call. Rather, typically in coming to know *that* I am waiting in my office, I will come to know *why* I am waiting in my office. Not always, but typically. So I come to know the proposition that I am waiting in my office because I am expecting a phone call, without previously knowing that I am waiting in my office. But I do not have sufficient evidence for the hypothesis that I am waiting in my office because I am expecting a phone call if I do not know that I am waiting in my office. A crucial piece of evidence is not in place until *after* I come to know why I am waiting in my office.

So we do not seem to have sufficient reasons for believing particular explanations about our own actions and attitudes, at least *prior* to knowing those explanations. There is a further and related consideration that suggests that we are not making an inference anyway. The point is that when we do explicitly consider the question of why we are doing what we are doing or think what we think, we do not consider facts which are reasons for and against believing some particular explanation or another of our action or attitude. Rather, we consider facts which are reasons for and against the action or the attitude itself. As Wittgenstein writes: "Asked: 'Are you going to do such-and-such'? I consider grounds for and against" (Wittgenstein, 1980, §815). And, as Stuart Hampshire writes:

If I am asked, “What do you intend to do?,” and if I were at all uncertain about the answer, I would normally consider reasons for acting in one way rather than another; that is, I would consider the merits of the various courses of action open to me. If I am asked “What do you believe?,” and if I were at all uncertain about the answer, I would normally consider the evidence in support of one proposition rather than another. (Hampshire, 1975, p. 59)

This is not something we would expect if we were arriving at knowledge of particular explanations of our actions and attitudes by inference. When I explicitly consider the question of whether there is life on Mars I consider reasons for and against believing that there is life on Mars. When I consider the question of whether I *believe* that there is life on Mars, I do *not* consider reasons for or against believing that I *believe* that there is life on Mars. I either consider no reasons at all, or I consider reasons for or against believing that there is life on Mars. Much has been made of this kind of observation in recent philosophical work (Evans, 1982a; Gallois, 1996; Moran, 2001; Byrne, 2005; Fernandez, 2013). Of course, one might argue that sometimes, in order to know what I believe, I first have to work out what to believe, and so will consider reasons for and against believing the thing in question before inferring that I believe it. But then we should expect one to consider reasons for and against believing that one believes the thing in question just after considering reasons for and against believing the thing in question. But this is not what we find.

A final related consideration is that often, when we know something by inference, we can say what the reasons are on the basis of which we believe what we do. Not always, of course, but often. Why do I believe that Wagner is a better composer than Verdi? Because Wagner had a greater harmonic range. Why do I believe that I believe that Wagner is a better composer than Verdi? Blank stare. There is no temptation to say ‘Well, I believe that Wagner is a better composer than Verdi, I believe that Wagner had a greater harmonic range than Verdi, so I figure I believe that Wagner is a better composer than Verdi because I believe that Wagner had a greater harmonic range than Verdi.’ Of course, there may be an alternative explanation of why we often do not know why we believe what we do about our own actions and attitudes which supports the inferential hypothesis. But absent such an explanation, this

is a consideration against that hypothesis.

So there are at least three considerations which strongly suggest that we do not know particular explanations of our actions and attitudes by inference, contrary to the inferential hypothesis. It is the existence of considerations like these which make the inferential hypothesis so hard to believe. And it is considerations like these which motivate theorists like Davidson to claim that such knowledge is not inferential. Of course, the considerations are not decisive. But Nisbett and Wilson do nothing to show that the inferential hypothesis can explain them or is consistent with them.

4.4 The Tension

At this point, then, we face a serious tension between the argument from symmetry and these considerations. Perhaps we should reconsider tentatively accepting the evidence for symmetry itself. But doing so on the basis of considerations like these would be a rather immodest application of common sense against empirical science. Perhaps we should reconsider the assumption, which we have tacitly accepted throughout, that observer subjects are arriving at explanations of the actions and attitudes of actor subjects by inference. Perhaps they are arriving at explanations by a kind of simulation rather than an inference. (See: (Heal, 1995, p. 47; Gordon, 1995, p. 71; Goldman, 1995, p. 78). For criticism of the simulation theory see: (Stich and Nichols, 1995, p. 123)). Perhaps. But it is not clear how that would help. Most simulation theorists accept that some kind of inference is involved. The question is whether simulation forms part of the process and replaces the role played by knowledge of a theory hypothesised on the theory-theory (Davies and Stone, 1995a,b). Perhaps, however, there is an alternative, and better, explanation of symmetry on which our knowledge is non-inferential, and has the features discussed in the previous section. That's what I am going to suggest in the next section.

But before doing so, let me discuss an alternative reaction. It might be thought that the argument from inadequate evidence given above doesn't show that our knowledge of our explanations cannot be inferential. Perhaps it just goes to show that if it is inferential, then the beliefs aren't based on sufficient reasons, and so aren't knowledge after all. This would be

to adopt a view on which our explanations of our actions and attitudes are confabulatory in the sense at play when Goldman writes that “people self-ascribe by a process of confabulation—the fabrication of ill-supported explanatory stories” (Goldman, 2006, p. 231). The key here is “ill-supported”. This would make Nisbett and Wilson’s argument more ambitious than it first seems. It is one thing to argue that our knowledge isn’t arrived at the way we think it is. It is another thing to argue that we don’t have knowledge where we appear to have. But it is a mistake to interpret Nisbett and Wilson in this way. First, Nisbett and Wilson explicitly say that “[t]he present analysis corresponds to common sense in that it allows that we will often be right about the causes of our judgements and behavior” (Nisbett and Wilson, 1977, p. 253). Since Nisbett and Wilson obviously want to make a concession to common sense here, it is important not only that we are right, in the sense that we have correctly guessed, but that we often *know* our reasons. Second, they go on in the second half of the paper to argue that in answering why questions we are making an inference to the best explanation, in light of the evidence available and background theory (Nisbett and Wilson, 1977, pp. 248–249). So Nisbett and Wilson are not endorsing a sceptical conclusion here.

4.5 An Alternative Explanation

As we have seen, much of the plausibility of Nisbett and Wilson’s argument derives from the fact that they assume that there are only two relevant alternative hypotheses concerning how we know particular explanations of our own actions and attitudes, the inner-sense hypothesis and the inferential hypothesis. This is a reasonable assumption to make, since there are no obvious alternatives to this hypothesis. However, in recent years, a kind of theory of self-knowledge has emerged as a rival to both inferential and inner-sense theories of self-knowledge. We can call this kind of theory an interrogative theory of self-knowledge, for reasons which will become apparent in a moment. In this section, I am going to sketch such an interrogative theory of self-knowledge and offer some initial motivation for the view. In the next section, I will argue that it offers an alternative explanations of the self and other symmetry.¹⁰

¹⁰The following discussion draws on (Hampshire and Hart, 1958; Hampshire, 1975, 1982) Similar ideas can be found in (Anscombe, 1963), but (Hampshire, 1975) and (Hampshire and Hart, 1958) emphasise the role of

4.5.1 *The Practical Question Theory*

The kind of alternative theory I want to examine here is motivated by the observation we noted above. The observation was that in answering theoretical questions about our own actions and attitudes we often consider reasons for and against performing the actions in question or for and against taking the attitudes in question. This observation tells against the inferential hypothesis, because according to the inferential hypothesis we would expect that in answering theoretical questions about our own actions and attitudes we would consider reasons for and against *believing* that we are performing particular actions or have taken particular attitudes. This observation is often made about non-explanatory theoretical questions about our own actions and attitudes like the question of whether we are doing this or that or the question of whether we believe this or that. But the observation is just as good in the case of explanatory theoretical questions, like the question of why we are doing what we are doing or the question of why we believe what we do. In these cases, we consider reasons for and against doing the thing in question or reasons for or against taking the attitude in question. We do not consider reasons for or against believing that particular explanations of our actions or attitudes obtain.

This observation motivates many alternative theories to the inferential and inner-sense theories. One theory which is particularly well motivated by this observation holds that we can arrive at knowledge about our own actions and attitudes by resolving particular practical questions about our own actions and mental states. For example, according to this theory, I can resolve the question of whether I will have another beer by resolving the question of whether to have another beer. Since I resolve the question of whether I will have another beer by resolving the question of whether to have another beer I do not consider reasons for and against believing that I will have another beer. Rather, I consider reasons for and against having another beer. Resolving the question of whether to have another beer involves considering reasons for and against having another beer, so if I resolve the question of whether I will have another beer by resolving the question of whether to have another beer I consider reasons for and against having another beer. If I resolve the question of whether to have another beer in the affirmative, then I decide to have another beer. If I resolve the question of

decision in non-inferential knowledge of action. Related ideas are developed in (Moran, 1988, 2001, 2012).

whether I will have another beer by resolving the question of whether to have another beer, then my belief that I will have another beer will be based on my decision to have another beer.

In order to distinguish this theory from alternatives, we can draw attention to the fact that it holds that one can resolve a *theoretical* question about one's own actions or attitudes, like the question of whether one *will* have another beer, by resolving a *practical* question about one's own actions or attitudes, like the question of whether *to* have another beer. The thought that we can resolve the theoretical question of what we will do by resolving the practical question of what to do is a natural one. And so is the thought that our beliefs about what we will do are based on decisions to do those very things. Such knowledge is, arguably, non-inferential, since it is directly based on our decisions to act, and not inferred from any else, not even facts about those decisions themselves. To resolve a theoretical question by resolving a practical question is not to arrive at a belief by inference.

Other theories have been motivated by the observation above. Some of these theories are alternative interrogative theories. They differ from the practical question theory in that they hold that we can resolve particular theoretical questions about our actions and attitudes by resolving other theoretical questions about our actions and attitudes, perhaps normative questions. Other theories downplay the idea of resolving one question by resolving another. I have chosen to focus on the practical question theory, because of all of these theories it provides the most promising alternative to the inferential and inner-sense theories.

4.5.2 Extending the Practical Question Theory

How might the practical question theory be extended to explain how we know particular explanations of our own actions and attitudes? As we just noted, when we consider explanatory questions about our own actions and attitudes we often consider reasons for and against performing those actions and for taking those attitudes. This is surprising, since, if the inferential theory were true, one would expect us to consider reasons for and against *believing* that particular explanations of our actions and attitudes obtain. But this is not what we find. Instead we find that we simply consider reasons for and against performing those actions and for and

and against taking those attitudes.

It is important to emphasise another negative claim: we do not consider reasons for and against *believing* that particular facts *are* reasons for or against performing some action or taking some attitude. We saw earlier that one way we might arrive at explanations by inference is from the fact that some fact is a good reason for performing some action, and the fact that some agent *has* that reason for performing that action. But this observation suggests that we consider *only* the reasons for and against acting themselves, and not reasons for and against believing that they *are* reasons for acting.

The practical question theory is well placed to explain this fact. It holds that we resolve theoretical questions about our own actions and attitudes by resolving practical questions about our own actions and attitudes. Since resolving practical questions about our own actions and attitudes involves considering reasons for and against performing those actions and taking those attitudes, resolving explanatory questions about our own actions and attitudes by resolving practical questions about our own actions and attitudes involves considering reasons for and against performing those actions and taking those attitudes.

But how exactly is the practical question theory to be extended to explain how we know particular explanations of our own actions and attitudes? So far we have just said that it holds that we can resolve *explanatory* questions about our own actions and attitudes by resolving practical questions about our own actions and attitudes. But what does that mean in this context? We have so far considered examples of resolving questions like the question of whether I will go to the party by resolving questions like the question of whether to go to the party. In such examples there is a nice symmetry between resolving each question in the affirmative and in the negative. One resolves the former in the affirmative if one resolves the latter in the affirmative, and resolves the former in the negative if one resolves the latter in the negative. But an explanatory question is far more discriminating.

At this point it is helpful to consider the case of resolving the question of *what* one will do by resolving the question of what *to* do. Often when one considers the questions of what one will do and what to do there is a determinate range of options one is considering. The options in the theoretical question will correspond to those in the practical question. The

theoretical question covers a range of things one might do, and so too does the practical question. Because of this symmetry one can resolve the theoretical question of what one will do by resolving the practical question of what to do. One's answer to the question of what one will do will correspond to one's answer to the question of what to do.

Now, there are two ways of developing the practical question theory in the case of explanatory questions. According to the first way, we fine-grain the content of the practical question in the way we do in the case of the question of what to do. That is, we make the practical question have options which correspond to the different answers to the theoretical question. For example, we might hold that one can resolve the question of whether one will go to the party because Mary is there by resolving the question of whether to go to the party because Mary is there, where 'whether to go to the party because Mary is there' is taken together. What is strange about this view is that it makes the reasons for which one acts part of the object of one's decision. On this view one decides *to go to the party because Mary is there*. This is not particularly plausible.

According to an alternative option, we do not fine-grain the content of the practical question in this way. Rather, we fine-grain the *ways* of resolving it. One might resolve it for one reason or for another. These are different *ways* of resolving the question. On this view, one might resolve the question of whether to go to the party because Mary is there, where 'because Mary is there' specifies the way one resolves the question. On this view, one decides to go to the party *because* Mary is there. I am not suggesting that the phrase 'decide to go to the party because Mary is there' concerns anything but an explanation of your decision. But I am suggesting that it corresponds to a different way of resolving the question of whether to go to the party. The relevant explanation obtains *because* one resolves the question in one way rather than another. It seems to me that resolving questions for different reasons corresponds to different *ways* of resolving the question, and these will correspond to different answers to the explanatory question.

One nice consequence of this view is that it explains why this would be a reliable way of forming beliefs about the explanations of our actions and attitudes. If we suppose that resolving practical questions for particular reasons is reliably connected with *doing* the action

in question for those reasons or *taking* the attitudes in question for those reasons, then there will be a reliable connection between the truth of one's belief and what it is about. For example, suppose I resolve the question of why I will go to the party by resolving the question of whether to go to the party in the affirmative for the reason that Jane is there. I believe that I will go to the party because Jane is there, and I *will* go to the party because Jane is there. If we assume, then, that resolving practical questions *in particular ways* is at the heart of the correct theory of what it is to do things for particular reasons or take particular attitudes for particular reasons, then we will have a very reliable way of forming beliefs about the explanations of our own actions and attitudes. However, insofar as there is a gap between resolving a practical question in a particular way and acting for a particular reason or taking a particular attitude for a particular reason, it will not be an infallible way of forming beliefs about the explanations of our own actions and attitudes.

Another nice consequence is that it promises to provide a way of responding to a long-standing problem for the practical question theory. That is the problem of robust resolutions. Suppose that I have already resolved the question of whether to go to the party but haven't resolved the question of whether I will go to the party. It seems that I can't come to know whether I will go to the party by resolving the question of whether to go to the party. I have already resolved that question. I cannot resolve it again. We can respond to this objection by pointing out that even though the question is already resolved *in a particular way*, we might resolve it *in a different way*. I might still resolve the question in the affirmative, but in a different way. So, even though I have already resolved the question of whether to go to the party, I might still consider the question of whether to go to the party and resolve the question in a different way and come to know that I will go to the party and why I will go to the party.

4.5.3 *Simulation and the Practical Question Theory*

The practical question theory receives further support from the fact that it provides a theory of how we can know what we *would* do in particular circumstances in a non-inferential way. Sometime our knowledge of what we *would* do in a particular circumstance is arrived at by inference from evidence we have about ourselves. But at other times it isn't. Rather, we

seem to imagine ourselves into the alternative circumstances and work out what we would do. There is a striking similarity between this case and the case of ordinary self-knowledge, for when we imagine ourselves into the alternative circumstance we consider reasons for and against particular actions in that circumstance. This suggests that the practical question theory can be applied here. Perhaps we resolve questions about what we *would* do in such circumstances by simulating resolving the question of what *to* do in those circumstances. We seem to have a way of running the process involved in the practical question theory offline. Moreover, it seems that we could use this process to work out not only what we will do but what we would do.

This observation has an important consequence. It provides observers with a way of coming to know what actors will do and why which is unlike ordinary theoretical inference. Rather, observers can come to know what actors will do and why by imagining themselves into the circumstances of the actor and resolving particular practical questions offline. They can then make an inference from what they would do and why in the particular circumstances to what the actor in question would do in those circumstances. This has important consequences for the explanation of symmetry, since if observers are using this method, it would explain why actors and observers often arrive at similar explanations. But as I am about to argue, even without this assumption, the practical question theory provides an alternative explanation of the self and other symmetries.

4.5.4 Explaining Symmetry

I have now introduced an alternative explanation of how we know particular explanations of our own actions and attitudes, which was motivated by some of the facts we examined earlier. The question now is whether this explanation provides a better explanation than the inferential hypothesis of the self and other symmetry. In this section, I will argue that it does.

What symmetry claims, is that there is a correlation between states, namely the actor subjects' beliefs about particular explanations, and the observer subjects' beliefs about particular explanations. Since it is unlikely that it is merely a coincidence that we find the pattern we do, and the actor subjects' beliefs about particular explanations are not a cause of the ob-

server subjects' beliefs about those explanations, nor vice versa, we should expect there to be a common cause of the actor subjects' and the observer subjects' beliefs (Reichenbach, 1971, p. 157). According to the inference hypothesis, the common cause of the actor subjects' and the observer subjects' beliefs are particular facts about the actor subjects' situation, including facts about the actor subjects' attitudes. Both actor subjects and observer subjects are assumed to be aware of these facts and this awareness is assumed to be caused by the obtaining of the facts themselves. (Of course, this relation doesn't hold in general between beliefs and what they are about. But it often does). The inferential theory can explain the symmetry because it provides an obvious mechanism by which both actor subjects' and observer subjects' beliefs can have a common cause.

But we are now in a position to see that there is an alternative explanation which can provide a mechanism by which both actor subjects' and observer subjects' beliefs can have a common cause. According to this theory, the actor subjects' beliefs and the observer subjects' beliefs have the same common cause hypothesised by the inferential theory: facts about the actor subject's attitudes. But whereas, in the case of observer subjects, the mechanism by which those facts cause the subject's beliefs in particular explanations goes via the subject's beliefs in those facts, in the case of the actor subjects, the mechanism does not go via the subject's beliefs in those facts. So the theory posits the same common cause, but a different mechanism by which they cause the agents' beliefs. The question now is why the different mechanisms give rise to the same pattern of error and ignorance.

The answer is this: the actor subject considers the contents of particular attitudes and aims to reach a rational decision on the basis of those contents on the basis of which she forms a belief concerning a particular explanation and the observer subject considers facts about those particular attitudes and aims to reach a belief about what rational decision the actor subject will make based on those contents and a belief about the explanation of the subject's response on this basis. If the observer subject is even moderately successful, we should expect her beliefs to coincide with the beliefs arrived at by the actor subject. Moreover, and most importantly, we should expect the kind of bias towards rationality we see in both the actors subject and observer subjects' explanations given these methods, as both methods will

be insensitive to non-rational factors. In this way our alternative theory can provide an explanation of symmetry. And it can do so on the assumption that observer subjects are making a theoretical inference to the best explanation concerning the actor subjects' responses.

To see how this explanation works, let's consider a case in detail. Suppose I consider the question of why I will go to the party, and I haven't previously considered the question. I consider reasons for and against going to the party, and resolve the question of whether I will go to the party on the basis of particular reasons. Suppose that one of those reasons is the fact that Jane is at the party, and that this is a particularly good reason for me to go to the party, as she is my best friend. Then I will come to believe that I am going to the party because Jane is there by resolving the question of whether to go to the party in the affirmative because Jane is there. I do not infer that this is why I am going to the party. Now, an observer who knew that the fact that Jane is at the party is a good reason for me to go to the party, and knew that I knew that Jane was at the party, would reasonably infer that I am going to the party because Jane is there. We arrive at the same hypothesis because the observer aims to give the best explanation of my action on the assumption that I am rational, and I am to *decide* to do things on the basis of the reasons I have. If I based my beliefs on my deciding to do something for a good reason, then we will arrive at similar explanations.

We are now in a position to see that the practical question theory provides an alternative explanation of how we know particular explanations of our actions and attitudes, one which explains why we can have such knowledge in the absence of sufficient reasons, explains why we attend to reasons for and against the actions and attitudes in question rather than reasons for and against believing that we are performing those actions or have those attitudes, and, finally, explains the observed self and other symmetries.

The question remains as to whether the practical question theory is the *best* explanation of these facts about our knowledge of rational explanations. One might object that it cannot be, since you cannot know something unless you have good reasons for believing it, whether or not you arrive at the relevant belief by inference or not. Thus, while our beliefs about our immediate environment are not arrived at by inference, we nonetheless have adequate reasons for believing the relevant propositions about our immediate environment. But this objection is

based on an implausible epistemological picture. The point about non-inferential knowledge is not only that it is not arrived at by inference, but that it needn't be justified in the way that inferential knowledge is. We needn't have adequate reasons for believing propositions arrived at in some non-inferential way. It is implausible to think that we have adequate reasons for believing basic propositions about our immediate environment which we come to believe on the basis of perception.

One might concede this point and then hold that you typically cannot know some contingent proposition unless you arrive at your belief by inference or on the basis of perceptual experience, whether or not you have adequate reasons for believing that proposition. I suspect that it is an epistemological view like this which motivates the thought that the inferential theory and the inner-perception theory are the only potential explanations, at least among some theorists. But this objection too rests on an epistemological prejudice. Knowledge of our future actions is neither based on inference or perceptual experience. So there is no reason to think that, in general, you cannot know contingent propositions unless you know them by inference or on the basis of perceptual experience.

More needs to be done in explicating the practical question theory and defending it against objections. But I think that we are at least in a position to see that it is a plausible explanation of the facts about our knowledge of rational explanations of our actions and attitudes. Unlike the inferential theory and the inner-perception theory, it does not seem to face any serious objections.

4.6 Conclusion

The claim that we each have a non-inferential way of knowing rational explanations of our own actions and attitudes, has a diminished status in contemporary theorising about introspection, compared to the claim that we each have a non-inferential way of knowing our own actions and attitudes themselves.¹¹ Many theorists happily concede our knowledge of rational explanations to the inferential theory, while holding that doing so is not at odds with the

¹¹The case of non-inferential knowledge of our actions is more controversial than the case of non-inferential knowledge of our attitudes. It has its defenders (Anscombe, 1963; Hampshire and Hart, 1958; Setiya, 2007, 2008, 2009a) as well as its opponents (Paul, 2009b,a)

claim that we each have a non-inferential way of knowing our own actions and attitudes. The following passage from Brie Gertler is representative of this position:

Even the staunchest proponents of privileged access acknowledge that we lack privileged access to these causal relations. So we should be wary of attempts to challenge the general idea of privileged access by citing cases in which subjects are ignorant of the causal sources of their attitudes or actions to challenge the general idea of privileged access. (Gertler, 2011, p. 75)¹²

If what I have said in this essay is correct, then this attitude is misguided. We have much to learn by taking sceptical challenges to non-inferential knowledge of particular explanations of our own actions and attitudes seriously.

4.7 Appendix: Further Evidence

4.7.1 Split Brain Studies

In this appendix I examine further empirical evidence which bears on the question of how we know particular explanations of our own actions and attitudes.

The first piece of empirical evidence comes from the split-brain studies of Michael Gazzaniga and his colleagues (Gazzaniga and LeDoux, 1978). The evidence concerns a single split-brain subject's answers to why-questions about his current actions. The particular split-brain subject was unique among split-brain patients in that he maintained significant linguistic abilities in his right hemisphere. Unlike other split-brain patients "... his right hemisphere can spell, and in addition, it can comprehend verbal commands, as well as process other parts of speech and make conceptual judgements involving verbal information" (Gazzaniga and LeDoux, 1978, p. 145). This meant that the subject was "... able to acting in response to verbal commands exclusively presented to either hemisphere but can describe verbally only the left-hemisphere stimuli" (Gazzaniga and LeDoux, 1978, p. 146).

¹²She goes on to say: "... [T]o show that we lack privileged access to such matters is to attack a straw man. ... [T]here are independent reasons to doubt that we have privileged access to the causal powers or etiologies of our mental states" (Gertler, 2011, p. 75). The independent reasons are *a priori* reasons for thinking we cannot know causes non-inferentially or at least that we cannot know them by observation. By "privileged access" Gertler means an epistemically privileged, and special way of knowing.

In an experimental setting, Gazzaniga and LeDoux prompted the subject to perform particular actions via the right hemisphere and then asked the subject to explain why he was doing what he was doing. The subject did not hesitate to answer. For example, in one experiment, the subject was prompted to laugh by being presented with the word “laugh” to the right hemisphere. When asked to explain why he was laughing, the subject replied “Oh, you guys are really something” (Gazzaniga and LeDoux, 1978, p. 146). This reply was consistent with a more general fact based on other experiments involving the subject, namely, that “the left hemisphere’s verbal response was consistent with the information available externally, but inconsistent with the true state of affairs only known by the right hemisphere” (Gazzaniga and LeDoux, 1978, pp. 147–148). According to Gazzaniga and LeDoux, when the subject was asked “Why are you doing that?” “...his talking left hemisphere was faced with the cognitive problem of explaining a discrete overt movement of great clarity carried out for reasons truly unknown to it” (Gazzaniga and LeDoux, 1978, p. 146). And it is a task that the left hemisphere performed well, without the subject being suspicious that there was anything abnormal going on.

In another experiment, Gazzaniga and LeDoux simultaneously presented the subject with a snow scene to the left hemisphere and a picture of a chicken’s claw to the right hemisphere and asked the subject to choose a card which best related to the object depicted from a choice of four for the left hand and four for the right hand. The subject selected a card depicting a snow shovel with his left hand and a card depicting a chicken with his right hand. The subject was then asked what he saw and why he chose what he did, to which he answered “I saw a claw and I picked the chicken, and you have to clean out the chicken shed with a shovel” (Gazzaniga and LeDoux, 1978, p. 148). Here, the left hand, controlled by the right hemisphere had chosen the shovel in response to the picture of the snow scene. However, the left hemisphere, which was not aware of the snow scene, swiftly found an explanation for why the shovel had been chosen. As Gazzaniga and LeDoux put it: “[t]he left hemisphere could easily and accurately identify why it had picked the answer, and then subsequently, and without batting an eye, it would incorporate the right hemisphere’s response into the framework” (Gazzaniga and LeDoux, 1978, pp. 148–149). Moreover, these suggestions were

not offered "...in a guessing vein but rather as a statement of fact as to why that card had been picked" (Gazzaniga and LeDoux, 1978, p. 149). In later work Gazzaniga posits an "interpreter" which monitors and synthesises the activities of other modular systems and it located in the left hemisphere in order to explain this ability (Gazzaniga, 1995, p 1393).

Finally, Gazzaniga offers the following case:

In still another example, if the command "walk" is flashed to the right hemisphere, the patient will typically stand up from the chair and begin to take leave from the testing van. When asked where he or she is going, the subject's left brain says, "I'm going into the house to get a Coke." However this test is manipulated, it always yields the same kind of result. (Gazzaniga, 1995, p 1393)

In light of their findings, Gazzaniga and LeDoux asked whether they were "... observing a basic mental mechanism common to us all" (Gazzaniga and LeDoux, 1978, p. 149). Gazzaniga and LeDoux's research is often taken as evidence for the view that beliefs concerning psychological explanations are confabulations, that is beliefs arrived at by confabulation. Here is how Timothy Wilson summarises their research:

...Gazzaniga and LeDoux have made the startling suggestion that we all share the tendency to confabulate, arguing that the conscious verbal self often does not know why we do what we do and thus creates an explanation that makes the most sense.... It may seem a substantial leap to conclude, on the basis of a few patients with brain damage or surgical sections, that all humans are blind to the causes of their actions and therefore have a "confabulator" that invents reasons. Yet there are times when the abilities and deficits of brain-damaged people provide a window into what it is like to be human, in addition to showing that some abilities are lost when the brain is damaged. Gazzaniga and LeDoux had the insight that severing the connection between the hemispheres might not have caused the kinds of confabulations they observed in [their subject]; rather, it made it easier to see a common human tendency to confabulate. (Wilson, 2002, p. 97)

Talk of confabulation here suggests that the means by which the subject is arriving at explanations of his choice is epistemologically suspect. It is a kind of *bad* inference or ill-supported inference. If this is so, then Gazzaniga and LeDoux's conclusion is the rather radical one that we typically do not *know* psychological explanations concerning our own attitudes and actions, since they are arrived at by faulty inferences. Gazzaniga himself puts his view in terms of "interpretation" rather than "confabulation", and the latter does not have the negative epistemological connotations. Of course, like Nisbett and Wilson, Gazzaniga and LeDoux do not think that we never *know* why we do what we do, as this passage suggests. This is evident when they write that "[t]he left hemisphere could easily and accurately identify why it had picked the answer" (Gazzaniga and LeDoux, 1978, p. 148). It is rather that in arriving at this explanation, the left hemisphere is drawing on "information available externally."

We might, then, take talk of "interpretation" to mean that the subject is arriving at his explanation by a kind of inference. And this is suggested by Wilson's claim that "... the conscious verbal self often does not know why we do what we do and thus creates an explanation that makes the most sense". After all, Gazzaniga thinks that the "interpreter" is located in the left hemisphere, and also that the "human inferential system" is "limited to the left hemisphere" (Gazzaniga, 1995, p. 1397). The conclusion that the subject is arriving at his belief by inference is almost irresistible, then.

It is helpful at this point, however, to recall the observation that we tend to consider reasons for and against a particular response and not reasons for and against believing that we have made that response. Here, in coming to a belief about why he chose the items that he did, the split-brain subject is looking to facts, or what he takes to be facts, which are normative reasons *for* the choice in question, and *not* normative reasons for or against *believing* some explanation or another of his choice. Gazzaniga and LeDoux all but make this point when they notice that "the left hemisphere's verbal response was consistent with the information available externally, but inconsistent with the true state of affairs only known by the right hemisphere." If the subject *were* making an ordinary inference, then we would expect him to look to facts about his own psychological states, and facts about what it is rational to believe in light of those mental states. This observation tells strongly *against* the inferential view, or

at least the view that the split-brain subject is performing an ordinary inference.

Now, it seems to me that Gazzaniga and LeDoux are providing direct evidence for the view that in the first-person case, we arrive at explanations by considering reasons *for* and *against* the relevant responses and resolving particular practical questions. The split-brain subject straightforwardly moves from such reasons to explanations of his own responses. It is implausible to think that the split-brain subject is making an inference from assumptions *about* his beliefs and mental states, or about whether the fact is, in fact, a reason for the response. Rather, he simply seems able to offer facts which *are*, from his point of view, reasons for the response, as part of an explanation of his response. This is what we should expect if he were arriving at his belief by resolving a practical question.

4.7.2 *The Choice Blindness Paradigm*

The next piece of evidence comes from the “choice blindness” studies of Johansson et. al (2005, 2006). Johansson et. al. explicitly see their choice blindness studies as reinvigorating the tradition begun by Nisbett and Wilson (Johansson et al., 2006, pp. 673–675). They note that although a few studies in the eighties tried to improve on Nisbett and Wilson’s methodology, the debate came to a standstill by the end of the decade after the publication of the original essay (Johansson et al., 2006, p. 674). And they go on to say that “[a]fter almost thirty years of intensive research on human cognition, it really *ought* to be possible to improve on the experimental design of Nisbett and Willson (1977)” (Johansson et al., 2006, p. 675). Let’s look at their attempt to do so.

In their original experiment Johansson et al. (2005) showed pairs of pictures of female faces to participants in the experiment and “...asked them to choose which face they found most attractive” (Johansson et al., 2005, p. 117). “On some trials, immediately after their choice, they were asked to verbally describe the reasons for choosing the way they did. Unknown to the participants on certain trials, a double card ploy was used to covertly exchange one face for another” (Johansson et al., 2005, p. 117). They go on to report that “We found no differences in the number of empty reports (when participants were unable to present any reasons at all) or in the degree to which reports were phrased in the present or past tense

(which might indicate whether the report is made in response to the present face or the prior context of choice” (Johansson et al., 2005, pp. 117–118). Moreover, in some cases the reason given could not have accounted for the original choice: “I chose her [the blond woman] because she had black hair” (Johansson et al., 2005, p. 118). In other cases the reason given could not apply to the face presented: “I chose her because she smiled [said of the solemn one]” (Johansson et al., 2005, p. 118).

In their discussion of their results Johansson et al. (2005) claim that these experiments “show, unequivocally, that normal participants may produce confabulatory reports when asked to describe the reasons behind their choices” (Johansson et al., 2005, p. 119). On their interpretation, then, we can imagine the subject making a quick inference to the the best explanation given the information available. Thus, when asked why he had chosen the woman with the black hair, the participant is currently aware that the woman has black hair and infers that, in light of this, he chose her because she had black hair. Likewise, it is plausible that in the other case, when the participant is asked why he chose the woman with the solemn face, he is remembering the fact that she was smiling and inferring that this best explains why he chose her. In a follow-up study, Johansson et al. (2006), used objective measures to determine that there was no relevant difference between the answers given in cases where the subject’s choices were switched. Further studies have even suggested that one will offer reasons in favour of a political or moral decision which one did not make (Johansson et al., 2012).

So we seem to have evidence for “confabulation” in non-clinical, relatively ordinary, settings. It’s worth pausing for a moment, however, and asking just what the inference is that participants are supposed to be making here. This isn’t something these researchers go in for, but it is worth asking. How exactly does our subject go from noticing that the woman in the photo has black hair to believing that he chose her because she had black hair? The evidence suggests that the participant is starting out from the premise that the woman has black hair. Perhaps the inference is something like this: this woman has black hair and I am aware that she had black hair, it is reasonable evaluate someone in terms of their attractiveness on the basis of the colour of their hair, therefore, I chose her because she has black hair.

In his discussion of these studies Goldman writes:

There is a different possible explanation for these findings, however. It seems likely that a participant is commonly attending to the photograph currently being perceived and forming new reasons on the fly, even though the conversation is *supposed* to concern his reasons for the choice made a moment earlier. . . .If this is right, the subject isn't literally confabulating; he is reporting respects of attractiveness genuinely operative in his current thinking. He may be introspecting these respects. (Goldman, 2006, p. 234)

Goldman's alternative explanation sounds very much like the practical question theory. To say that the subject is 'forming new reasons on the fly' is to say he is resolving the practical question *in a different way*, for different reasons. The subject isn't merely inferring that some explanation or another obtains, he is arriving at an explanation by resolving a practical question. So, it seems that the evidence from the choice blindness studies supports the practical question theory over the inferential theory.

CHAPTER 5

INTERROGATIVE THEORIES OF INTROSPECTION

5.1 Introduction

I don't know whether I will have another beer. I haven't yet resolved the question of whether I will have another beer. I also don't know whether *to* have another beer. I haven't yet resolved the question of whether *to* have another beer. In my situation, I can, it seems, resolve the *theoretical question* of whether I will have another beer—a straightforward question of fact about the future—in one of two ways, either by considering reasons for and against *believing* that I will have another beer, or by resolving the *practical question* of whether to have another beer. Indeed, it seems that I can come to know whether I will have another beer by resolving the practical question of whether *to* have another beer. If I do, then, my knowledge will, arguably, be non-inferential. Alternatively, I might come to know whether I will have another beer by considering reasons for and against *believing* that I will have another beer, and inferring that I will from such reasons. If I do, then my knowledge will be inferential.

In general, it seems that we resolve *theoretical* questions about whether we will do something in particular, and so come to know whether we will do it, in a non-inferential way,

by resolving corresponding *practical questions* about whether *to do* the thing in question. One might be motivated by this observation to defend a kind of *interrogative theory* of non-inferential knowledge of action.¹ An interrogative theory of non-inferential knowledge of action holds that we can resolve *theoretical* questions about whether we will do something in particular, and so come to know whether we will do it, in a non-inferential way, by resolving some related question, like the question of whether *to do* the thing in question, or whether one *should* do the thing in question.

The primary aim of this essay is to explicate and distinguish three central kinds of interrogative theory of introspection, the *practical question theory*, the *theoretical question theory*, and the *normative question theory*, and to examine the prospects of each theory as a general theory of how we can come to have non-inferential knowledge of our actions and attitudes in general. I argue that the practical question theory is the most promising kind of interrogative theory of introspection, and a very plausible, but often misconstrued and misunderstood, theory of how we can come to have non-inferential knowledge of our own action and attitudes.

One important challenge for the *practical question theory* is to explain how we can come to have non-inferential knowledge of whether we believe some particular proposition. According to the practical question theory, I can resolve the theoretical question of whether I *believe* Jane will have another beer, and so come to have non-inferential knowledge of whether I *believe* that Jane will have another beer, by resolving the question of whether *to believe* that Jane will have another beer. It seems however, that I can only resolve the question of whether *to believe* that Jane will have another beer either by *deciding* to believe that Jane will have another beer or by *deciding* not to believe that Jane will have another beer. But it is implausible to think that I can come to know whether I believe that Jane will have another beer by *deciding* to believe that she will or *deciding* not to believe that she will. Thus, it is thought, by both proponents and critics of the theory, that in the case of belief, resolving the

¹The basic idea behind this theory of knowledge of action is in Wittgenstein (Wittgenstein, 1980). Wittgenstein doesn't put the point in terms of resolving one question by resolving another, but rather in terms of considering reasons for and against doing the thing in question in contrast to considering reasons for and against believing that one will do it. The initial development of the basic idea into a theory is in (Hampshire, 1975) and (Hampshire and Hart, 1958).

question of whether *to* believe that Jane will have another beer is either the same thing as resolving the theoretical question of *whether Jane will have another beer* or the same thing as resolving the theoretical question of *whether I should believe that Jane will have another beer*.² In other words, in the case of belief, the practical question theory is either implausible, since it requires one to decide to believe, or is equivalent to either the normative question theory or the theoretical question theory.

One of the central claims of this essay is that this is not the case. Resolving the question of whether to believe that Jane will have another beer does not involve deciding to believe that Jane will have another beer or deciding not to believe that she will have another beer, and nor is it the same thing as resolving the question of whether Jane will have another beer, or resolving the question of whether I should believe that Jane will have another beer. This has important consequences for the prospects of the practical question theory as a general theory of how we can come to have non-inferential knowledge of our actions and attitudes.

Here is how I will proceed. In the next section, §2, I introduce three *prima facie* distinct interrogative theories of introspection: the *practical question theory*, the *theoretical question theory*, and the *normative question theory*. In the next two sections, §3 and §4, I introduce an analytic framework for thinking about interrogative theories of introspection and argue that the three theories are in fact distinct. In §3, I argue that the intuitive distinction between practical and theoretical questions, which I have relied on in these introductory remarks, corresponds to a semantic distinction between the denotation of particular readings of particular interrogative clauses. I argue that, whereas interrogatives like ‘whether I will have another beer’ denote sets of propositions which are their possible answers, interrogatives like ‘whether to have another beer’ denote sets of possible events which are the alternative options associated with the interrogative. In §4, I argue that the practical question theory is distinct from the normative question theory in general: resolving the question of whether *to* do something is not the same thing as resolving the question of whether one *should* do it, in general.

²A prominent defender of the former view is Richard Moran (Moran, 2001). The latter view is often attributed to Moran by his critics. As I understand him, Moran defends a version of the practical question theory in general, but is tempted by the theoretical question theory in the case of belief since he seems to think that resolving the question of whether to believe that Jane will have another beer is the same thing as resolving the question of whether Jane will have another beer.

In §5, I argue that the practical question theory is distinct from the theoretical question theory in the case of belief: resolving the question of whether *to* believe that something is the case is not the same thing as resolving the question of whether something is the case.

Having argued that the three interrogative theories introduced in §2 are in fact distinct, I evaluate each theory in the following three sections, that is, in §§6-9 respectively. I ask whether any of these theories provide a plausible unified theory of how we can come to have non-inferential knowledge of facts about our actions and attitudes, including positive, negative, and explanatory facts. Each theory is evaluated in terms of two important criteria: *reliability* and *rationality*. I ask whether the theory provides a *reliable* way of forming beliefs about our actions or attitudes in various cases, and whether the theory plausibly provides a *rational* way of forming beliefs about our actions or attitudes in various cases. In §6 I evaluate the normative question theory. I argue that in no case does it provide a reliable or a rational way of forming beliefs about our actions or attitudes. Moreover, it faces what I call the *mismatch problem*: it fails to provide a way of knowing that we do *not* believe some proposition when we do not believe the proposition but do not believe the negation of that proposition. In §7 I evaluate the theoretical question theory. I argue that although the theory provides a highly reliable way of forming beliefs about our own beliefs, it fails to provide a reliable way of forming beliefs about our actions and other attitudes. Moreover, it also faces the mismatch problem. Finally, in §8 I evaluate the practical question theory. I argue that once we are clear about the theory, and do not confuse it with the normative question theory or the theoretical question theory in particular cases, it provides a suitably reliable and rational way of forming beliefs about our own actions and attitudes, including negative beliefs and explanatory beliefs, across the board, and does not face the mismatch problem. I conclude that the practical question theory is the most promising kind of interrogative theory of introspection, and is a very plausible theory of how we can come to have non-inferential knowledge of our own action actions and attitudes in general.

5.2 Interrogative Theories of Introspection

Interrogative theories of introspection hold that we can come to know answers to theoretical questions about our own actions and attitudes, like the question of whether we believe some particular proposition, by resolving related but distinct questions, like the normative question of whether we *should* believe some proposition, or the practical question of whether *to* believe some proposition. Several contemporary theories of introspection can be thought of as interrogative theories of introspection, although they are not so-called by their proponents.³ Particular attributions raise interpretative issues which we will not examine here. Instead, I propose to explicate several distinct introspective theories beginning from first principles.

We can distinguish three salient kinds of interrogative theory according to the following intuitive distinction between three different kinds of questions:

(1) Practical Questions

E.g. whether to have another beer, (controversially: whether to believe that it is raining).

(2) (Non-Normative) Theoretical Questions

E.g. whether it is raining, whether I believe that it is raining.

(3) Normative (Theoretical) Questions

E.g. whether I should have another beer, whether I should believe that it is raining.

The three salient kinds of interrogative theory, are, then:

(1) The Practical Question Theory (PQT): one resolves theoretical questions about one's actions and attitudes by resolving *practical questions* about one's actions or attitudes.

³Two particularly salient theories are those given in (Evans, 1982b; Moran, 2001).

(2) The Theoretical Question Theory (TQT): one resolves theoretical questions about one's actions and attitudes by resolving related (non-normative) *theoretical questions*.

(3) The Normative Question Theory (NQT): one resolves theoretical questions about one's actions and attitudes by resolving *normative* (theoretical) *questions* about one's actions and attitudes.

A few examples might help to get clear on the distinction between these theories. According to the PQT, I can resolve the question of whether I believe that Jane will have another beer, say, by resolving the practical question of whether *to* believe that Jane will have another beer. According to the TQT, I can resolve the question of whether I believe that Jane will have another beer, to keep with the same example, by resolving the *theoretical question* of whether Jane will have another beer. According to the NQT, by contrast, I can resolve the question of whether I believe that Jane will have another beer by resolving the *normative* theoretical question of whether I *should* believe that Jane will have another beer.

What each theory has in common with the others is that they hold that one can resolve one question by resolving a distinct but related question. I assume that the idea is intuitive enough, although a couple of examples will help. I might resolve the question of whether Jane will have another beer by resolving the question of whether Jane *wants* another beer. This is a familiar case where I might resolve one *theoretical* question by resolving another *theoretical* question. In this case, resolving one question by resolving another involves *inferring* one proposition—the proposition that Jane will have another beer—from another proposition—the proposition that Jane wants another beer. I might resolve the question of whether to take the train to Hannover by resolving the theoretical question of whether I must take the train to Hannover in order to get to Hamburg. This is a familiar case where I might resolve a *practical* question by resolving a *theoretical* question. In this case I do not infer one proposition from another. Deciding to take the train to Hannover on the basis of the fact that I must take the train to Hannover in order to get to Hamburg is not, on the face of it, a matter of inferring one proposition from another (although one might want to speak of a 'practical inference' here).

On the face of it, the three theories we have just distinguished are three distinct theories

of introspection, and, as such, might have importantly different consequences. But it might be thought that the theories collapse into each other in particular cases. In particular, it might be thought that the PTQ collapses into the NQT either across the board or at least in the case of belief. That is, it might be thought that resolving the question of whether *to* have another beer is not distinct from resolving the question of whether one *should* have another beer, or it might be thought that resolving the question of whether *to* believe that Jane will have another beer is not distinct from resolving the question of whether one *should believe* that Jane will have another beer. Moreover, it might be thought that the PTQ collapses into the TQT in the case of belief, since it might be thought that resolving the practical question of whether to believe that Jane will have another beer is not distinct from resolving the theoretical question of whether Jane will have another beer.

In the next three sections I argue against the view that these views are not distinct. In the next section, §3, I argue that theoretical and practical questions are distinct kinds of questions. In §§4-5 I argue that resolving a practical question is *always* distinct from resolving a theoretical question. It follows that the PQT is distinct from the NQT and the TQT in every case.

5.3 Practical Questions and Theoretical Questions

What is the distinction between practical questions like the question of whether to have another beer and theoretical questions like the question of whether I will have another beer or the question of whether I should have another beer? It cannot be that practical questions are about actions whereas theoretical questions are not, since the theoretical questions are about actions in the relevant sense. In this section I argue that the difference between theoretical questions and practical questions corresponds to a difference in the *kind* of semantic object denoted by different kinds of interrogatives. More precisely I argue that theoretical questions are, or at least determine, partitions of possible worlds (sets of possible worlds propositions), the cells of which correspond to propositions which are possible answers to the question, whereas practical questions are, or at least determine, partitions of possible events (sets of possible events), the cells of which correspond to sets of possible events which are the *op-*

tions associated with the question.

It is tempting, at least initially, to draw the distinction between theoretical questions and practical questions in terms of the distinction between the denotation of finite interrogative clauses like ‘whether I will have another beer’ and the denotation of infinitival interrogative clauses like ‘whether to have another beer’. On this way of drawing the distinction, finite interrogative clauses denote theoretical questions and infinitival interrogative clauses denote practical questions.

Finite interrogatives like ‘whether I will have another beer’ denote sets of propositions which correspond to possible answers to the question or partitions of possible worlds the cells of which correspond to possible answers to the question (Hamblin, 1958; Karttunen, 1977; Groenendijk and Stokhof, 1982; Higginbotham, 1993). For example, the finite interrogative ‘whether Jane will have another beer’ denotes a bi-partition of possible worlds the cells of which correspond to the proposition that Jane will have another beer and the proposition that Jane will not have another beer. So we can think of theoretical questions as being partitions of possible worlds, or at least determining partitions of possible worlds. Then we might simply claim that whatever the denotation of infinitival interrogatives is, it is not a partition of possible worlds.

The trouble with drawing the distinction between theoretical and practical questions in terms of the denotation of finite and infinitival interrogatives is that, while there is widespread agreement about the denotation of finite interrogatives, there is little agreement about the denotation of infinitival interrogatives. Some theorists think that the infinitival interrogative ‘whether to have another beer’ denotes a theoretical question, that is, it denotes a partition of possible worlds. Moreover, it has been widely noted that infinitival interrogatives have readings on which they are equivalent to finite interrogatives on particular readings (Huddleston and Pullum, 2005; Stanley and Williamson, 2001; Stanley, 2011). Consider:

(1a) Jane wondered whether to have another beer.

(1b) Jane wondered whether she should have another beer.

(1a) seems to have a reading on which it is equivalent to (1b). Since the embedded interrogative in (1b) denotes a bi-partition of possible worlds the cells of which correspond to

the proposition that Jane should have another beer and the proposition that Jane should not have another beer, it seems that the embedded interrogative in (1a) must denote the same bipartition of possible worlds.

So if we are going to draw the distinction between theoretical questions and practical questions in terms of the denotation of finite and infinitival interrogatives, we cannot do so in such a straightforward way.

Nonetheless, I think that it can be done. Although infinitival interrogatives have readings on which they denote theoretical questions and are equivalent to particular finite interrogatives, they also seem to have readings on which they are *not* equivalent to any finite interrogative. If so, then we can draw the distinction between practical and theoretical questions in terms of the denotation of infinitival interrogatives *on our favoured reading* and the denotation of finite interrogatives in general.

Suppose that we are watching Jane. We see her look at her empty glass, and then towards the bar. Now consider our pair of sentences again:

(1a) Jane (just) wondered *whether to have another beer*.

(1b) Jane (just) wondered *whether she should have another beer*.

Suppose Jane knows that she shouldn't have another beer. Knowing this, she might nonetheless wonder whether *to* have another beer. Then we are tempted to think that (1a) is true, and that (1b) is false. Indeed, we are tempted to think that there is no reading of (1b) on which it is true. Jane did not wonder whether she *should* have another beer in any sense of 'should'. Nonetheless, she wondered whether *to* have another beer. She can consider her options without wondering whether she *should* do something or other. In general, it seems that we can know that we shouldn't do something in every sense of 'should', and still wonder whether *to* do it.

Suppose that Jane has decided not to have another beer. Having decided, she might nonetheless wonder whether she *should* have another beer. Then we are tempted to think that (1b) is true and (1a) is false. Indeed, we are tempted to think that Jane might wonder whether she *should* have another beer, in any sense of 'should' and still not wonder whether

to have another beer. She can wonder whether she *should* do something without reconsidering her decision.

So, (1a) can be true on some reading while (1b) is false on every reading, and (1b) can be true on some reading while (1a) is false on some reading. The best explanation of this fact is that the embedded interrogative in (1a) does not have the same denotation as the embedded interrogative in (1b) on the relevant readings. (An alternative explanation is that ‘wonder’ is ambiguous between infinitival and finite interrogatives. But that isn’t a very plausible explanation, consider: ‘Jane wondered whether to have another beer and whether she had enough money to pay for it’). Indeed, it is on this reading that we feel that the embedded interrogative in (1a) does not denote a theoretical question, but, rather, denotes a practical question.

What is a practical question, then? We can think of a practical question as a partition of possible events. Then the infinitival interrogative ‘whether to have another beer’ denotes a partition of possible events the cells of which correspond to a set of possible events of having another beer and a set of possible events of not having another beer. These sets of possible events are, intuitively, the agent’s *options*. On this view, theoretical questions and practical questions are different kinds of *semantic* objects.

One might think the considerations above do not show that infinitival interrogatives do not have readings on which they denote practical questions, understood in this way, rather than denoting theoretical questions. Just because the finite interrogative ‘whether she should have another beer’ does not have a reading on which it is equivalent to ‘whether to have another beer’ it does not follow that ‘whether to have another beer’ does not denote a theoretical question. Perhaps it denotes a theoretical question which is not denoted by ‘whether she should have another beer’ on any reading. Of course it doesn’t follow. But the argument isn’t meant to be deductive. The argument is that the best explanation of the non-equivalence here is that on this reading the infinitival interrogative does not denote the same kind of semantic object as the finite interrogative. The point is that if ‘whether to have another beer’ does not denote a bi-partition of possible worlds which can be denoted by ‘whether she should have another beer’ on some reading, then this is because it doesn’t denote a bi-partition of possible worlds at all.

Although I have argued that theoretical and practical questions are distinct kinds of semantic objects, let me hedge a little for the argument to follow. In the next section I am going to argue that resolving practical questions is distinct from resolving theoretical questions of all kinds. If the reader disagrees with me about the distinction I have drawn, I am still able to make my point in the following way: my claim should be that resolving this special kind of theoretical question is distinct from resolving ordinary theoretical questions. In order to distinguish between the PQT and the NQT I will then need to argue that the PQT involves this special kind of theoretical question and the normative question theory involves an ordinary theoretical question. This would suffice for my purposes.

5.4 Resolving Practical and Theoretical Questions

According to interrogative theories of introspection we can resolve theoretical questions about our own mental states by resolving distinct but related questions. The notions of resolving a question, and resolving one question by resolving another are central to interrogative theories of introspection. What is it to resolve a question? And what is it to resolve one question by resolving another question? (In speaking of *resolving* a question rather than *answering* or *settling* I am simply making an arbitrary choice among what I take to be equivalent terms for the psychological notion I am interested in).

We have a pre-theoretical understanding of the notion of resolving a question. I don't know whether to go to the cafe this morning. I am still thinking about whether to go. Finally, I resolve the question of whether to go to the cafe this morning. I decide to go. To resolve the question of whether to go to the cafe this morning is, it seems, either to decide to go to the cafe or to decide not to go to the cafe. I don't know whether Hamburg is west of Berlin. I think about whether it is. I consult a map and resolve the question of whether Hamburg is west of Berlin. I come to believe that it is. To resolve the question of whether Hamburg is west of Berlin is, it seems, either to come to believe that Hamburg is west of Berlin or to come to believe that Hamburg is not west of Berlin. Since the question of whether to go to the cafe this morning is a practical question and the question of whether Hamburg is west of Berlin is a theoretical question, we might be tempted to adopt the following two theses about

resolving practical and theoretical questions on the basis of these examples:

Theoretical Resolution: to resolve a theoretical question is to come to believe an answer to the question.⁴

Practical Resolution: to resolve a practical question is to decide to take an option determined by the question.

On this view, there is a connection between the semantic notion of a theoretical question and a psychological claim about what it is to resolve a theoretical question, and there is a connection between the semantic notion of a practical question and a psychological claim about what it is to resolve a practical question.

What is it to resolve one question by resolving another question? Again, we have a pre-theoretical understanding of the notion of resolving one question by resolving another. I might resolve the question of whether to go to the cafe by resolving the question of whether I should go to the cafe. This is not to deny that to resolve the question of whether to go to the cafe is either to decide to go to the cafe or decide not to go to the cafe. I resolve the question of whether to go to the cafe only if I either decide to go to the cafe or decide not to go to the cafe. But I can decide to go to the cafe on the basis of my belief that I should go to the cafe. If I do, then I resolve the question of whether to go to the cafe by resolving the question of whether I should go to the cafe. Only confusion will result from not being clear about the distinction between what it is to resolve a question and what it is to resolve one question by resolving another.

Once we have these distinctions in mind, it becomes pretty clear that I can resolve the practical question of whether to go to the cafe without resolving the theoretical question of whether I should go to the cafe, and that I can resolve the theoretical question of whether I should go to the cafe without resolving the practical question of whether to go to the cafe. This follows if I can decide to go to the cafe without coming to believe that I should, and if I

⁴One might worry that it is possible to come to believe an answer to a theoretical question without resolving that question. Perhaps I believe that some theory is correct, but I continue to seek evidence for and against it. Then, it seems, the question of whether the theory is correct is not resolved. I am tempted to claim that in this case, it is the question of whether *to* believe that the theory is correct which remains unresolved, whereas the question of whether the theory is correct is resolved. Since nothing will rest on this claim, I will set it aside in what follows.

can come to believe that I should go to the cafe without deciding to go to the cafe. Similarly, I can resolve the practical question of whether to go to the cafe without resolving the theoretical question of whether I will go to the cafe, and I can resolve the question of whether I will go to the cafe without resolving the question of whether to go to the cafe. This might be difficult to see because I might often resolve the question of whether I will go to the cafe by resolving the question of whether to go to the cafe, as the PTQ holds. Then I will come to believe that I will go to the cafe on the basis of my decision to go to the cafe. But, again, this does not conflict with the claim that to resolve the question of whether I will go to the cafe is either to come to believe that I will go to the cafe or to come to believe that I will not go to the cafe.

5.4.1 Resolving Practical Questions About Attitudes

In the previous section we examined the thesis that to resolve a practical question is to decide to take an option determined by the question. This thesis was motivated by considering examples about action. But the thesis is problematic when it comes to what on the face of it appear to be practical questions about attitudes, like the question of whether to believe that Jane will have another beer, the question of whether to intend to have another beer, and the question of whether to want to travel to Australia. According to the thesis, to resolve the question of whether to believe that Jane will have another beer is either to decide to believe that Jane will have another beer or to decide not to believe that she will have another beer, and similarly for the other examples. There is something incredibly awkward about the claim that to resolve the question of whether to believe that Jane will have another beer is either to decide to believe that Jane will have another beer or to decide not to believe that she will have another beer. This has motivated theorists to adopt a different thesis about what it is to resolve a practical question about belief:

Practical/Theoretical Resolution: to resolve a practical question about belief is to resolve a theoretical question about the content of the belief.

Practical/Normative Resolution: to resolve a practical question about belief is to resolve a normative question about the belief.

For example, according to Practical/Theoretical Resolution, to resolve the practical question

of whether to believe that Jane will have another beer is to resolve the theoretical question of whether Jane will have another beer, and according to Practical/Normative Resolution, to resolve the practical question about whether to believe that Jane will have another beer is to resolve the normative (theoretical) question of whether one should believe that Jane will have another beer. If one holds Practical/Theoretical Resolution, then one holds that the PTQ is equivalent to the TQT in the case of belief. If one holds Practical/Normative Resolution, then one holds that the PTQ is equivalent to the NQT in the case of belief. One might be tempted to generalize from the case of belief to other attitudes, thus holding that the PTQ is equivalent to either the TQT or the NQT in the case of attitudes, but not the in the case of actions.

When it comes to resolving practical questions about attitudes, then, it seems that we must either hold the implausible view that we resolve these questions by deciding to take some option determined by the question, or hold one of the theses above. I think, however, that there is an alternative. We should deny that to resolve a practical question about our attitudes is to decide to take some option determined by the question, and we should deny both of the theses above. In what follows, I argue against both of the theses above. This will motivate the recognition of a genuine alternative here.

Consider the claim of Practical/Theoretical Resolution, that to resolve a practical question about belief is to resolve a theoretical question about the content of the belief. Consider, in particular, the claim that to resolve the question of whether to believe that Jane will have another beer is to resolve the question of whether Jane will have another beer. We know that to resolve the latter question is either to believe that Jane will have another beer, or to believe that Jane will not have another beer. In order to show that Practical/Theoretical Resolution is false, we can provide a case where one resolves the question of whether to believe that Jane will have another beer without coming to believe that Jane will have another beer or coming to believe that Jane will not have another beer. But this is exactly what happens when one withholds judgement about whether Jane will have another beer. One resolves the question of whether to believe that Jane will have another beer in the negative without resolving the question of whether Jane will have another beer. So, it is not the case that resolving the question of whether to believe that Jane will have another beer is either coming to believe

that Jane will have another beer or coming to believe that Jane will not have another beer.

But this doesn't yet show that if I believe that Jane will have another beer, or believe that she will not have another beer, then I have resolved the question of whether to believe that she will have another beer. But, it seems to me, there are such cases. Consider the father who cannot bring himself to believe that his son committed a crime. He considers reasons for and against believing that his son committed the crime, and resolves the question of whether to believe that his son committed the crime in the affirmative. But he doesn't believe that his son committed the crime. So, it seems to me, that we are entitled to the stronger conclusion that coming to believe some proposition is not sufficient for resolving the question of whether to believe that proposition.

It is cases like these which encourage the proponent of Practical/Normative Resolution. On this view, cases of withholding judgement are to be understood as cases where one resolves the question of whether one should believe some proposition in the negative, and cases like the father's are to be understood as cases where one resolves the question of whether one should believe some proposition in the affirmative even though one does not believe the proposition. If we accept Practical/Normative Resolution then we can explain why it is possible to resolve the question of whether to believe some proposition without believing that proposition or believing its negation. For the rest of this section I am going to argue that this is not the best explanation of these cases. Practical/Normative Resolution requires a level of sophistication which is simply not present in such cases.

Consider the case of withholding judgement again. Suppose I consider the question of whether to believe that Jane will have another beer. Suppose that I withhold judgement. I resolve the question of whether to believe that Jane will have another beer in the negative. According to Practical/Normative Resolution, to resolve the question of whether to believe that Jane will have another beer is to resolve the question of whether I should believe that Jane will have another beer. Now, suppose that we are playing a drinking game, and Jane's turn is not up yet. I am waiting to see whether she rolls a six. I am withholding judgement about whether Jane will have another beer until I see whether she rolls a six. I have resolved the question of whether I should believe that Jane will have another beer in the negative on

the basis of the fact that her turn in the drinking game is not up yet. That's my reason for resolving the question in the negative. It is a reason for resolving the question of whether to believe that she will have another beer in the negative, it is a reason for not believing that Jane will have another beer. But if to resolve the question of whether to believe that Jane will have another beer is to resolve the question of whether I should believe that Jane will have another beer, then I shouldn't consider reasons for and against believing that Jane will have another beer. I should consider reasons for and against believing that I should believe that Jane will have another beer. But the fact that Jane's turn in the drinking game is not up yet is not a reason for or against believing that I should believe that Jane will have another beer. Rather, the fact that the fact that Jane's turn in the drinking game is not up yet is a reason against believing that Jane will have another beer is a reason for believing that I should not believe that she will have another beer. But I am not plausibly aware of this fact, or any fact like it. To consider whether I should believe that Jane will have another beer is to consider reasons for and against believing that I should believe that Jane will have another beer. But I simply do not seem to do this. Rather, I consider reasons for and against believing that Jane will have another beer. This suggests that it is not the case that resolving the question of whether to believe that Jane will have another beer is to resolve the question of whether I should believe that Jane will have another beer.

A similar point can be made about the case of the father. In order for the father to resolve the question of whether to believe that his son committed the crime, according to Practical/Normative Resolution, the father must consider the question of whether he should believe that he son committed the crime, and not just the question of whether to believe that his son committed the crime. The former involves considering normative propositions about his belief, the latter involves considering options concerning belief. In order to resolve the question of whether he should believe that his son committed the crime, the father must consider reasons for and against believing that he should believe that his son committed the crime. This set of reasons is not coextensive with the reasons for and against believing that his son committed the crime. Indeed, in order to answer the former question, the father needs to consider the theoretical question of whether particular facts are reasons for or against believing that

he should believe that his son committed the crime. In contrast, in order to answer the latter question, the father need only consider reasons which are in fact reasons for and against believing that his son committed the crime. For example, consider the fact that his son's fingerprints were found at the scene. The father is aware of this fact. It is a reason for believing that his son committed the crime. He might resolve the question of whether to believe that his son committed the crime on the basis of this reason among others. But this fact isn't a reason for believing that he should believe that his son committed the crime. Rather, the fact that it is a reason for believing that his son committed the crime is a reason for believing that he should believe that his son committed the crime. The father may not be aware of this fact at all.

These points suggest that we need to recognise that resolving the question of whether to believe some proposition is distinct from resolving the question of whether the proposition is true (which is to either to come to believe the proposition or to come to believe its negation) and is distinct from resolving the question of whether one should believe the proposition. The case of withholding judgement, and cases where one has resolved the question of whether to believe some proposition, but does not believe it or its negation, tell against the first option. I have argued that these cases should not be thought of as cases where one has resolved the normative question of whether one should believe the proposition in question. Resolving the normative question of whether one should believe the proposition in question requires a level of sophistication which goes well beyond what we intuitively think is present in the relevant cases.

An important consequence of this conclusion is that resolving the question of whether to believe some proposition stands to believing the proposition as resolving the question of whether to do something stands to doing it, even though to resolve the question of whether to believe something is not to decide to believe it. Resolving the question of whether to believe some proposition is not just to come to believe that proposition or believe its negation, just as resolving the question of whether to do something is not just either to do the thing or not to do the thing. Resolving the question of whether to believe some proposition is not just to come to believe that one should believe that proposition or that one should not believe that proposition,

just as resolving the question of whether to do something is not just to come to believe that one should do that thing, or that one should not do that thing. Nonetheless, if one resolves the question of whether to believe some proposition, then, typically, one believes in accordance with how one has resolved the question, just as if one resolves the question of whether to do something, typically, one acts in accordance with how one has resolved the question. This can be seen most clearly in the case where one already believes some proposition, considers the question of whether to believe it, and withholds judgement. In such a case one resolves the question of whether to believe the proposition in the negative. This should, rationally speaking, have the consequence that one ceases to believe the proposition in question.

This completes my case for distinguishing between the PTQ, the NQT, and the TQT. I have argued that even in the case of the attitudes, these theories are all distinct.

5.5 Evaluating the Theoretical Question Theory

In this section and the next two I evaluate the TQT, the NQT, and the PQT respectively. The following two criteria will be central to the evaluation:

Reliability: a way of forming beliefs about whether P is a way of coming to know whether P only if it is reliable.

Rationality: a way of forming beliefs about whether P is a way of coming to know whether P only if it is rational.

I will assume that Reliability is an uncontroversial necessary condition for knowledge. I will not assume that Rationality is, although I will nonetheless take it to be a mark against a theory that it does not seem to provide a rational way of forming beliefs.

In the case of belief, the TQT holds:

Belief: one resolves the question of whether one believes that P by resolving the theoretical question of whether P.

This claim is perhaps the most widely discussed claim of interrogative theories of introspection (Evans, 1982b; Moran, 2001; Shah and Velleman, 2005). On this view, the question of

whether one believes that P is said to be ‘transparent’ to the question of whether P. Moreover, it is said to be ‘transparent’ to the world, since the theoretical question of whether P typically is not about one and one’s beliefs but is about whatever one’s belief is about. (Both of these claims are emphasized by Moran (Moran, 2001, Ch. 2)).

It has been noted that this would be a very reliable way of arriving at beliefs about one’s beliefs (Byrne, 2005, pp. 96–98). To see why, notice that if one resolves the question of whether one believes that P in the affirmative by resolving the question of whether P in the affirmative, then one’s belief that one believes that P will be true. Likewise if one resolves the question of whether one believes that P in the negative by resolving the question of whether P in the negative. So the TQT as applied to the case of belief meets the reliability criterion.

Unfortunately, it is not obvious how one might generalise the TQT beyond the case of belief, since belief provides a special case where a theoretical question about whether one believes a particular proposition can be resolved by resolving a question about that proposition. There are no strict analogues for intention, desire, action, and forms of explanation.⁵ It is for this reason that proponents of the TQT opt for something like the NQT for the case of intention, desire, and other attitudes more generally. Indeed, this fact leads some to adopt the NQT *rather than* the TQT across the board (Boyle, 2009, p. 136).

Nonetheless, suggestions have been made for the case of intention and explanation on behalf of the TQT which retain its spirit. One suggestion in the case of intention is the following:

Intention(2): one resolves the question of whether one intends to V by resolving (in a particular way) the question of whether one will V (Byrne, 2011, p. 216).

The case of intention requires a slight modification of the theory (indicated in parentheses). Suppose the theory held that I can resolve the question of whether I intend to have another beer by resolving the question of whether I will have another beer. Suppose I resolve the question of whether I will have another beer by considering reasons for and against believing

⁵The point is made in (Bar-On, 2004, pp. 114–115). This is noted by Cassam (Cassam, 2014, p. 4 & p. 103). He calls it the “Generality Problem for Rationalism” (Cassam, 2014, p. 103). Following a suggestion of David Finkelstein (Finkelstein, 2003, p. 103), Cassam suggests that the solution to the problem is to adopt something analogous to the NQT. See (Byrne, 2005, p. 100) for a sketch of a proposal for desire.

that I will have another beer. Then it seems that there is no connection between resolving this question and intending to have another beer.

We need a qualification: I must resolve the question of whether I will have another beer in a particular way, it seems. We might hold that I must resolve the question of whether I will have another beer by considering reasons for and against *having another beer*. If I resolve the question of whether I will have another beer in *this* way, then, arguably, I will intend to have another beer. So resolving the question of whether I intend to have another beer by resolving the question of whether I will have another beer is a reliable way of arriving at a true belief about whether I intend to have another beer. (The amendment is suggested by Byrne (Byrne, 2011, p. 218)).

This modification represents a departure from the TQT in its pure form. To resolve the question of whether one will have another beer in a *practical way* is to resolve the question of whether to V or to resolve the question of whether to intend to V. So one wonders whether the theory needs to appeal to the theoretical question of whether one will have another beer at all. Why not directly resolve the question of whether one intends to go to the party by resolving the question of whether to V or the question of whether to intend to V?

Intention(III): one resolves the theoretical question of whether one intends to V by resolving the practical question of whether to V.

Intention(IV): one resolves the theoretical question of whether one intends to V by resolving the practical question of whether to intend to V.

So the modified version of the TQT risks collapsing into a version of the PQT. (See (Baker, 2015) and (Paul, 2015) for related criticisms). A proponent of the TQT might resist this. But I think that resistance is futile. Even if the TQT can be defended for belief and intention, it simply cannot be generalised beyond these cases. As I have already said, in the case of desire, it collapses into the NQT, holding that the question of whether one wants to V is resolved by resolving the question of whether one should want to V, whether one should V, whether V-ing is desirable, whether V-ing is preferable, or some other normative question. The theory cannot be generalized to the case of action, and as the account of intention shows, it takes non-inferential knowledge of action for granted.

One interesting case for the TQT is the case of explanation of belief. In the case of explanation of belief, the following suggestion has been made (Setiya, 2013):

Explanation of Belief: one resolves the question of whether one believes that P because Q by resolving the questions of whether P and whether Q is a (good) reason to believe that P.

There is a sense in which this theory is a hybrid between the NQT and the TQT, since it involves both a (non-normative) theoretical question and a normative (theoretical) question. Nonetheless, it is broadly in the spirit of the TQT, so I will discuss it here.

The thought behind Explanation of Belief is that one can resolve the question, say, of whether one believes that Jane will have another beer because she has only had two, by resolving the question of whether Jane will have another beer and of whether the fact that Jane has only had two beers is a (good) reason for believing that Jane will have another beer. The theory has some *prima facie* plausibility.

However, the main problem with Explanation of Belief is that it is not a reliable way of forming beliefs about why one believes what one believes in general. Resolving the question of whether I believe that I was born in Australia because my birth certificate says so by resolving the questions of whether I was born in Australia and whether the fact that my birth certificate says so is a reason to believe that I was born in Australia is *not* a reliable way of forming a belief about whether I was born in Australia because my birth certificate says so.

Kieran Setiya disagrees (Setiya, 2013). Setiya holds that if I believe that I was born in Australia and believe that the fact that my birth certificate says so is a reason for me to believe that I was born in Australia then I believe that I was born in Australia because my birth certificate says so. Thus Setiya holds a theory of believing for a reason on which forming a belief in the way specified in Explanation of Belief would be highly reliable. This is a controversial theory of what it is to believe something for a reason, to say the least. If it can be defended, then perhaps this approach to knowledge of explanations of belief can be defended.

So the TQT seems to face serious problems with reliability once we move beyond the case of belief. But, it is not entirely unproblematic in the case of belief. Another problem for

the TQT is the *mismatch* problem.⁶ It accounts for only one of two cases where one resolves the question of whether one believes that P in the negative. One can resolve the question of whether one believes that P in the negative by resolving the question of whether P in the negative. But this will only account for the case where one does not believe that P because one believes that not P. It will not account for the case where one does not believe that P because one neither believes that P or does not believe that P. So not only is the theory unreliable when generalised beyond the case of belief, it cannot, it seems, be generalised to account for cases where we are obviously able to know particular propositions about our attitudes by introspection.

The TQT also faces a problem with rationality. If one resolves the question of whether one believes that P by resolving the question of whether P in the affirmative, then one's belief that one believes that P is based on one's belief that P *and nothing else*. It is as if one made a one premise inference from the proposition that P to the conclusion that one believes that P. Such an inference would be irrational. If your belief that you believe that P is only based on your belief that P then it appears to be irrational.

One way of responding to this objection is to downplay *rationality* as a criterion for knowledge, or to argue that this way of forming beliefs *is*, contrary to appearances, rational, or to ask the objector for their reasons for thinking that this way of forming beliefs is not rational. Notice that this objection is not that in order for a belief to be rational it must be based on sufficient reasons. I think that that thesis is mistaken (as I have argued in Ch. 2). Rather the objection is that if a belief is based on other beliefs then the contents of those beliefs had better be sufficient reasons for so believing.

Another way of responding is to draw attention to the difference between inferring that one believes that P from the proposition that P and resolving the question of whether one believes that P by resolving the theoretical question of whether P. One might insist that the process of *resolving* the question of whether P in resolving the question of whether one believes that P plays an important rational role, and makes this way of forming the belief rational. But why should it matter that the transition involve one's resolving the question of whether P?

⁶Objections of this kind to theories of introspection in general have been raised by (Sosa, 2003) and (Stoljar, 2012).

It seems to me that there is something right in this response however. If one infers that one believes that P from the premise that P, one's belief that one believes that P is irrational. But I think that it remains irrational even if one resolves the question of whether P in resolving the question of whether one believes that P. I will argue that the situation is improved if one resolves the question of whether one believes that P by resolving the question of whether *to* believe that P. In this case, one's belief that one believes that P is *not* based on one's belief that P: it is simply based on one's resolving the question of whether *to* believe that P. This way of forming beliefs about one's own beliefs is rational, in contrast to that of resolving the question of whether one believes that P by resolving the question of whether P.

In sum, the TQT faces problems concerning *reliability* and *rationality* in many cases, and also faces the mismatch problem.

5.6 Evaluating the Normative Question Theory

In the case of belief the NQT holds:

Belief: one can resolve the question of whether one believes that P by resolving the question of whether one *should* believe that P.

For example, I can resolve the question of whether Jane will go to the party by resolving the question of whether *I should* believe that Jane will go to the party. Much depends on the sense of 'should' involved in the theory. I will assume that the intended sense of 'should' is the *rational* sense of 'should'.

Arriving at a belief about whether one believes that P by resolving the question of whether one *should* believe that P would be a reliable way of forming true beliefs about whether one believes that P only if it were true that one tends to believe what one believes one rationally should believe.

One problem with the NQT is that there is a gap between what one believes one rationally should believe and what one believes (Cassam, 2014, p. 106–110). It may well be true that it is irrational to believe that you rationally should believe something and yet not believe it. But we are not perfectly rational. So this way of forming beliefs about our own beliefs will often not be reliable (Paul, 2012, pp. 334–335).

Another problem with the NQT is that in many cases we simply do not have beliefs about what we *should* believe. It would have us answer a relatively simple question by answering a rather complex question.⁷

In the case of intention the NQT holds:

Intention(1): one can resolve the question of whether one intends to V by resolving the question of whether one *should* V.

Intention(2): one can resolve the question of whether one intends to V by resolving the question of whether one *should intend* to V.

Intention(1) is not strictly analogous with Belief, but Intention(2) is. Again it has been argued that this is not a reliable way of forming true beliefs about whether one intends to V (Cassam, 2014). It is too easy for one to resolve the question of whether one intends to have another beer by resolving the question of whether one *should* have another beer, or whether one *should intend* to have another beer, and for one's belief about whether one intends to have another beer to be false.

An additional problem with Intention(2) is that we often simply do not have beliefs about what we *should* intend when we know what we intend to do by introspection. This objection also applies to Intention(1) but in a milder form. We often do not have beliefs about what we *should* do when we know what we intend to do by introspection. I intend to have another coffee. I know I do in a non-inferential way. I do not believe either that I should or should not have another coffee.

In the case of desire, the NQT holds:

Desire(1): one can resolve the question of whether one wants to V by resolving the question of whether one *should* V.

Desire(2): one can resolve the question of whether one wants to V by resolving the question of whether one *should want* to V.⁸

⁷This point is made forcefully by Cassam. Cassam calls it the "Substitution Problem for Rationalism" (Cassam, 2014, p. 104). A related objection is raised by (Finkelstein, 2003, p. 163). The problem is that we are not often *required* to have particular mental states, but rather are only *permitted* to have them. The point is also made by (Way, 2007, p. 228).

⁸Lawlor attributes this view to Moran. She attributes to Moran the claim that "... ordinarily, if one is rational,

Again, Desire(1) is not strictly analogous with Belief, but Desire(2) is. The NQT as applied to desire has been widely criticised (Ashwell, 2013). Resolving the question of whether one wants to V by resolving the question of whether one should want to V is a manifestly unreliable way of forming beliefs about whether one wants to V.

Let's turn to the case of action. The NQT holds:

Action(1): one can resolve the question of whether one will V by resolving the question of whether one should V.

Action(2): one can resolve the question of whether one is V-ing by resolving the question of whether one should be V-ing.

These are manifestly unreliable ways of forming beliefs about whether one will V and whether one is V-ing. I know of no defence of the NQT as applied to action.

The problem transfers to the case of explanation too:

Explanation of Belief(1): one can resolve the question of whether one believes that P because Q by resolving the question of whether one should believe that P because Q.

Explanation of Action(1): one can resolve the question of whether one will V because P by resolving the question of whether one should V because P.

But one might modify the NQT slightly.

Explanation of Belief(2): one can resolve the question of whether one believes that P because Q, given that one knows whether one believes that P, by resolving the question of whether one should believe that P because Q.

Explanation of Action(2): one can resolve the question of whether one will V because P, given that one knows whether one will V, by resolving the question of whether one should V because P.

a *descriptive* question about what one's attitude is, in fact presents a *deliberative* question about what one's attitude should be, in light of one's reasons (Lawlor, 2009, p. 53).

The thought here is that we assume that one knows whether one will V, then one can resolve the question of whether one will V because P by resolving the question of whether one should V because P. But even if this is granted, this is an unreliable way of forming beliefs about whether one will V because P.

Now let's consider *rationality*. According to the NQT, I can resolve the question of whether I will have another beer by resolving the question of whether I should have another beer. If I do, then my belief that I will have another beer is based on my belief that I should have another beer *and nothing else*. It is as if I made a one premise inference from the proposition that I should have another beer to the conclusion that I will have another beer. It matters little whether we call this 'inference'. Such transitions between belief have struck many as irrational. Many have held that they are rational only if one's belief is based on a further belief. The inferences for the other cases and the additional assumptions are represented below:

P1 I should believe that P.

P2 If I should believe that P, then I believe that P.

C1 Therefore, I believe that P.

P1 I should (intend to) V.

P2 If I should (intend to) V, then I intend to V.

C1 Therefore, I intend to V.

P1 I should want to V.

P2 If I should want to V, then I want to V.

C1 Therefore, I want to V.

P1 I should V.

P2 If I should V, then I will V.

C1 Therefore, I will V.

P1 I should V because P.

P2 If I should V because P, then I will V because P.

C1 Therefore, I will V because P.

But if these other beliefs *are* involved, then the NQT arguably becomes a inferential theory of introspection, and one which faces many objections. One objection is that we simply do not believe the background assumption. The second is that even if we do, then such beliefs are false and irrational, so forming beliefs in this way cannot be a way of coming to know the relevant propositions. It is advisable not to defend the NQT in this form. But then it still faces the difficulty that the transition seems to be irrational. It is hard to see how to evaluate it except in terms of a single premise inference. Refusing to call it an inference doesn't do anything to overcome the problem that it seems irrational.

In sum, the NQT faces problems concerning *reliability* and *rationality* across the board, even in the case of belief. The problem with *reliability* cannot be overcome by idealising, since it is not ideal reliability that matters for knowledge, but *actual* reliability.

5.7 Evaluating the Practical Question Theory

In the case of belief, the PQT holds:

Belief: one resolves the theoretical question of whether one believes that P by resolving the practical question of whether to believe that P.

It seems to me that theorists have been blind to the possibility of the PQT being applied in the case of belief, and have thought that the TQT and the NQT, are the only alternative interrogative theories, because they have thought that resolving the question of whether to believe that P must either (i) involve deciding to believe that P (ii) involve believing that one should believe that P or (iii) involve believing that P. Since (i) is a manifestly implausible

option, it is assumed that (ii) or (iii) are the live options. But I argued above that resolving the question of whether to believe that P is not a matter of deciding to believe that P, or of believing that one should believe that P, or of believing that P. We must recognise an alternative. We must resist views which reduce resolving the question of whether to believe that P to deciding to believe that P, believing that one should believe that P, or believing that P. Resolving the question of whether to believe that P is not equivalent to any of these, although it can be helpfully thought of on analogy with decision.

Resolving the question of whether one believes that P by resolving the question of whether to believe that P is a reliable way of forming a belief about whether one believes that P. This is because one typically believes that P if one resolves the question of whether to believe that P in the affirmative, and does not believe that P if one resolves the question of whether to believe that P in the negative. The connection here is weaker than that posited by the TQT as applied to belief, but stronger than that posited by the NQT. Resolving the question of whether to believe that P in the affirmative stands to believing that P as resolving the question of whether V in the affirmative stands to V-ing. The relation is stronger than the relation between resolving the question of whether one *should* V and one's V-ing.

In the case of intention, desire, and action, the PQT holds:

Intention: one resolves the question of whether one intends to V by resolving the question of whether to intend to V.

Desire: one resolves the question of whether one wants to V by resolving the question of whether to want to V.

Action: one resolves the question of whether one will V by resolving the question of whether to V.

Of course, the question of whether to intend to V, whether to want to V, and whether to V, rarely come apart. Nonetheless, one can resolve the question of whether to V in the affirmative while resolving the question of whether to intend to V in the negative. For example, I might resolve the question of whether to breathe steadily in the affirmative and resolve the question of whether to intend to breathe steadily in the negative if I know that intending to breathe

steadily will have a negative effect on my breathing steadily.⁹

In the case of explanation of belief and action the PQT holds:

Explanation of Belief: one resolves the question of whether one believes that P because Q by resolving the question of whether to believe that P because Q.

Explanation of Action: one resolves the question of whether one will V because P by resolving the question of whether to V because P.

Suppose it looks to me like there is a yellow mug in front of me. I can resolve the question of whether I believe that there is a yellow mug in front of me because it looks to me like there is a yellow mug in front of me by resolving the question of whether to believe that there is a yellow mug in front of me because it looks to me like there is a yellow mug in front of me. If I know that I am hallucinating, then I may well resolve the question in the negative. But if I do not, then I may well resolve it in the positive. Whichever way I resolve the question will make a difference not only to whether I believe that there is a yellow mug in front of me, but also to whether or not I believe that there is a yellow mug in front of me *because* it looks like there is a yellow mug in front of me.

Since resolving questions about what to believe and what to do makes a difference to what one believes and why one does, the PQT appeals to a reliable way of forming beliefs about what one believes and what one will do. Since resolving questions about whether one believes something for some particular reason or whether one will do something for some particular reason makes a difference to whether one believes it for that reason, or whether one will do it for that reason, the PQT appeals to a reliable way of forming beliefs here.

The PQT does not face the *mismatch* problem. There is a match between resolving the question of whether to believe that P in the negative and resolving the question of whether one believes that P in the negative, and likewise for intention, desire, and explanation. That is, there is a correspondence between the alternative answers to the question of whether to believe that P and the question of whether one does believe that P which there isn't between the question of whether P and the question of whether one believes that P.

⁹This example is adopted from another context. See (Paul, 2009b; Setiya, 2009a).

Finally, the PQT avoids the criticism concerning *rationality*, since it does not posit a transition between beliefs. Take the case of action. In this case, in resolving the question of whether one will V by resolving the question of whether to V one's belief that one will V is based on one's *decision* to V alone, and not on any belief one has. Intuitively, this is a rational way of forming beliefs about what one will do. The PQT posits an analog of decision in the case of belief, and it plays a similar rational role.

In sum, the PQT can easily be generalized across the board. It does not face immediate objections from Reliability, Rationality, or the mismatch problem. If it is assumed that resolving the question of whether to believe that P is distinct from deciding to believe that P, believing that one should believe that P, and believing that P, and that resolving the question of whether to believe that P plays a psychological and rational role analogous to resolving the question of whether to V, the PQT is the most unified and plausible interrogative theory of introspection.

5.8 Two Problems for the Practical Question Theory

I now want to discuss two related *prima facie* problems for the PQT: the problem of fragile resolutions and the problem of robust resolutions. Both of these problems are also problems for the TQT, so they will offer a further opportunity to compare and contrast the theories.

5.8.1 *Fragile Resolutions*

Suppose I believe that my keys are on the table, so I have resolved the question of whether my keys are on the table in the affirmative. But suppose further that if I were to consider the question of whether my keys are on the table, or whether *to* believe that my keys are on the table, I would immediately cease to believe that they are on the table, and come to believe that they are on the bench. So my affirmative resolution of the question of whether my keys are on the table is *fragile* in the sense that my positive resolution will change to a negative resolution if I consider the question of whether my keys are on the table or whether to believe that my keys are on the table.

Suppose that I consider the question of whether I believe that my keys are on the table.

According to the PQT I can resolve this question by resolving the question of whether to believe that my keys are on the table. According to the TQT I can resolve this question by resolving the question of whether my keys are on the table. Both theories predict that I will resolve the question of whether I believe that my keys are on the table in the affirmative, since I will resolve the question of whether *to* believe that my keys are on the table in the affirmative, and I will resolve the question of whether the keys are on the table in the affirmative, respectively. If the belief I arrive at is about whether I believe that my keys are on the table *prior* to resolving the relevant questions, then my belief will be false. If the belief I arrive at is about whether I believe that my keys are on the table *posterior* to resolving the relevant questions, then my belief will be true. So, it seems that the PQT and the TQT provide a reliable way of knowing only whether one believes something *posterior* to resolving the relevant questions, and not *prior* to resolving the relevant questions.

Some theorists see an objection to the PTQ and the TQT in this observation (Shah and Velleman, 2005; Gertler, 2009, p. 506–508). The objection is simply that resolving the question of whether *to* believe that my keys are on the table would be an unreliable way resolving the question of whether I believe that my keys are on the table *at some time prior* to resolving the question of whether *to* believe that my keys are on the table. This point, of course, should be conceded (Byrne, 2011, p. 208 n. 10). It would be an unreliable way. If I want to know whether I believe that my keys are on the table at the time I consider the question of whether they are, then, since it takes time to resolve such a question, it seems that I must make an inference about whether I believed that my keys were on the table on the basis of reasons for believing that I believed that my keys were on the table. So if I want to find out what I believed a moment ago, I shouldn't do something which tells me what I believe now. But it is hard to see how this is an objection to the PQT.

A more pressing objection would hold that we have a non-inferential way of knowing such facts about the immediate past. One could then argue that since the TQT and the PQT cannot account for this, they are not fully general theories of non-inferential knowledge of such facts. Or, more carefully, one would then have an important consideration to bear in mind when trying to figure out what the best explanation of how we can know facts about our

own actions and attitudes is, since providing a more unified explanation is a mark in favor of one theory over another. Some theories of introspection hold that we *do* have a non-inferential way of knowing such facts about the immediate past. These theories model introspection on perception. Just as I can know on the basis of a perceptual memory that you were wearing a red shirt when I saw you today, I can know on the basis of an ‘inner-perceptual’ memory that I believed that my keys were on the table moments ago (Lyons, 1986). Needless to say, such theories are not particularly plausible. I am tempted to simply deny that we can know facts about the immediate past in a non-inferential way. When I know whether I believe that my keys are on the table in a non-inferential way, I know *de se* and *de nunc* that I believe that my keys are on the table.

There is one other way of filling out the objection. According to this version, I have a non-inferential way of resolving the question of whether I believe that my keys are on the table which will *not* make a difference to whether I believe that my keys are on the table. If so, the PTQ and the TQT cannot explain such knowledge, since they both predict that the way I have of knowing whether I believe that my keys are on the table *will* make a difference to whether I believe that my keys are on the table.

But it seems to me that in such a case I simply *do not* have a non-inferential way of knowing whether I believe that my keys are on the table which will *not* make a difference to what I believe. So the objection turns into an argument *for* the PTQ or the TQT and against alternative theories.

Some non-interrogative theories predict that we do have a way of knowing fragile resolutions. An inner-sense theory predicts that I should be able to know whether I believe that my keys are on the table in a way that will not make a difference to whether I believe that my keys are on the table. I should be able to bracket the question of whether *to* believe that my keys are on the table and bracket the question of whether the keys are on the table and simply resolve the question of whether I believe that my keys are on the table in some non-inferential way—i.e., on the basis of inner-perception. A special-inferential theory like Byrne’s predicts that I should be able to know whether I believe that my keys are on the table in a way that will not make a difference to whether I believe that my keys are on the table. According to this theory,

I simply make an inference from the premise that my keys are on the table to the conclusion that I believe that my keys are on the table (hence the label 'special-inferentialism'). I should be able to bracket the question of whether *to* believe that my keys are on the table and bracket the question of whether the keys are on the table and simply resolve the question of whether I believe that my keys are on the table in some non-inferential way—i.e., by inferring that I believe that my keys are on the table from the premise that my keys are on the table. Indeed, a version of the TQT theory which allows that I can resolve the question of whether I believe that my keys are on the table on the basis of my *having resolved* the question of whether my keys are on the table makes the same prediction, since I *have* resolved the question of whether my keys are on the table, I would just resolve it in a different way if I considered the question now. So the only theory, it seems, which makes the right prediction here is the PQT which holds that I must resolve the question of whether I believe that my keys are on the table by resolving the question of whether *to* believe that my keys are on the table.

5.8.2 Robust Resolutions

Suppose I believe that Hamburg is in Germany. I have resolved the question of whether Hamburg is in Germany. If I were to consider the question of whether Hamburg is in Germany, I wouldn't change my mind. Let's say that my resolution is *robust*. Suppose that I consider the question of whether I *believe* that Hamburg is in Germany. It seems that I can come to know that I believe that Hamburg is in Germany in a non-inferential way. Indeed, it seems that I can resolve the question of whether I believe that Hamburg is in Germany without resolving the question of whether Hamburg is in Germany, since I have already resolved the question, and my resolution is *robust*.

This raises a problem for the TQT, since it holds that I resolve the question of whether I *believe* that Hamburg is in Germany by resolving the question of whether Hamburg is in Germany (Shoemaker, 2003; Byrne, 2005, pp. 84–85; Byrne, 2011). Some theorists have appealed to this kind of case to reject interrogative theories of introspection like the TQT (Byrne, 2005, pp. 84–85). The objection certainly is a pressing one for the TQT. One might argue that, contrary to appearances, I *do* resolve the already resolved question of whether to

believe that Hamburg is in Germany in such cases. But it seems to me that if a question is already resolved, one cannot resolve it.

But how does the objection bear on the PQT? One might argue that in such cases, one does not resolve the question of whether *to* believe that Hamburg is in Germany. But this is an additional claim to the original one. There seem to be many cases where one resolves the question of whether one believes that P in a non-inferential way without resolving the question of whether P, but it isn't at all obvious that there are many cases where one resolves the question of whether one believes that P in a non-inferential way without resolving the question of whether *to* believe that P. As we just saw, the case of fragile resolutions gives us good reasons for thinking that we *must* resolve the question of whether we believe that P by resolving the question of whether *to* believe that P. So, taken together, the observation about robust resolutions and the observation about fragile resolutions support the PTQ over the TQT and neither observation is the basis for an objection to the PTQ. Of course, one might try to argue that we *can* resolve the question of whether we believe that P in a non-inferential way without resolving the question of whether *to* believe that P. But that case hasn't been made, and the case of robust resolutions does not support the case.

5.9 Conclusion

I began by distinguishing between three kinds of interrogative theory of introspection: the practical question theory, the normative question theory and the theoretical question theory. I argued that these theories were distinct because resolving normative questions, theoretical questions, and practical questions did not amount to the same thing. I then evaluated each theory. I have argued that the practical question theory is the best interrogative theory of introspection and that it is a plausible theory of how we know particular facts about our actions and attitudes, including, importantly, explanatory facts, in a non-inferential way. The framework developed here allowed us to see that many of the objections to the practical question theory rest on misunderstandings of the theory and on confusing it with other kinds of interrogative theory. The question remains, at the end of this essay, whether there is a more general theory of how we know our actions, attitudes, and *other mental states* available

and whether the practical question theory might be generalised to other mental states. Those questions will have to wait for another day.

CHAPTER 6

CONCLUSION

I have decided to pursue a few lines of argument in detail in each essay rather than attempting to pursue a single line of argument throughout the thesis. If the arguments are on the right track, then I think that they constitute a significant rethinking of the problem of self-knowledge. If my arguments about the connection between reasons and rationality are correct, then solutions to the problem of self-knowledge need not be constrained by the requirement that a belief is rational only if it is based on sufficient reasons for so believing. We are free to pursue solutions which do not meet this condition. If my arguments about our knowledge of particular explanations of our own actions and attitudes are correct, then we can see that a fully unified theory of introspection must provide an answer to the question of how we know particular explanatory facts about our own actions and attitudes, given that such knowledge is non-inferential. But such a solution should explain the observed symmetries between self and other. I argued that the practical question theory of introspection can do this. If my arguments concerning the practical question theory are correct, then theorists should take more care in distinguishing between interrogative theories of introspection. Moreover, if my arguments are correct, then the practical question theory of introspection may be more plausible than it has recently been taken to be. Much more needs to be done to develop such a theory, but I hope to have provided further motivation for doing so. I also hope to have clarified the distinction

between reason explanations and other kinds of explanations by defending a particular thesis about the semantics of 'because' sentences. Being clear on this distinction will allow us to better evaluate claims about how we come to know particular explanations of our own actions and attitudes.

BIBLIOGRAPHY

- Achinstein, Peter. 1983. *The Nature of Explanation*. Oxford University Press.
- Alvarez, Maria. 2010. *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford University Press.
- Anscombe, Elizabeth. 1963. *Intention*. Harvard University Press.
- Ashwell, Lauren. 2013. “Deep, dark ... or transparent? Knowing our desires.” *Philosophical Studies* 165:245–256.
- Bach, Kent. 1994. “Conversational Implicature.” *Mind & Language* 9.
- . 1999. “The Myth of Conventional Implicature.” *Linguistics and Philosophy* 22:pp. 327–366.
- . 2001. “You Don’t Say?” *Synthese* 128:15–44.
- Baker, Derek. 2015. “Why Transparency Undermines Economy.” *Synthese* .
- Bar-On, Dorit. 2004. *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Clarendon. ISBN 9780199263202.
- Beebe, Helen. 2003. “Seeing Causing.” *Proceedings of Aristotelian Society* 103.

- Bem, Daryl J. 1967. "Self-Perception: An Alternative Interpretation of Cognitive Dissonance Phenomena." *Psychological Review* 74:183–200.
- Boghossian, Paul. 1989. "Content and Self-Knowledge." *Philosophical Topics* XVII:5–26.
- Boyle, Matthew. 2009. "Two Kinds of Self-Knowledge." *Philosophy and Phenomenological Research* 78:133–164.
- Brewer, Bill. 1999. *Perception and Reason*. Oxford University Press.
- Byrne, Alex. 2005. "Introspection." *Philosophical Topics* 33:79–104.
- . 2011. "Transparency, Belief, Intention." In *Aristotelian Society Supplementary Volume*, volume 85, 201–221. Wiley Online Library.
- Carnie, Andrew. 2006. *Syntax: A Generative Introduction*. Blackwell Publishing.
- Carruthers, Peter. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.
- Cassam, Quassim. 2014. *Self-Knowledge for Humans*. Oxford University Press.
- Clifford, William K. 1877. "The Ethics of Belief." *Contemporary Review* 29:289.
- Cruse, D. A. 1986. *Lexical Semantics*. Cambridge University Press.
- Culicover, Peter W. and Jackendoff, Ray. 2005. *Simpler Syntax*. Oxford University Press.
- Cullity, Garrett and Gaut, Berys (eds.). 1997. *Ethics and Practical Reason*. Clarendon Press.
- Dancy, Jonathan. 2000. *Practical Reality*. Oxford University Press.
- Davidson, Donald. 1963. "Actions, Reasons, and Causes." In *Essays on Actions and Events*. Oxford University Press.
- . 1976. "Hempel on Explaining Action." In *Essays on Actions and Events*. Oxford University Press.
- . 1980. *Essays on Actions and Events*. Oxford: Oxford University Press.

- Davies, Martin and Stone, Tony (eds.). 1995a. *Folk Psychology*. Blackwell.
- . 1995b. *Mental Simulation*. Blackwell.
- Dray, William H. 1960. *Laws and Explanation in History*. Oxford University Press.
- Dretske, Fred. 1972. “Contrastive Statements.” *The Philosophical Review* 81:411–437.
- . 2011. “Skeptical Doubts About Self-Knowledge.” In Sven Bernecker and Duncan Pritchard (eds.), *The Routledge Companion to Epistemology*. Routledge.
- Ericsson, K. Anders and Simon, Herbert A. 1980. “Verbal Reports as Data.” *Psychological Review* 87.
- Evans, Gareth. 1982a. *Varieties of Reference*. Oxford University Press.
- . 1982b. *The Varieties of Reference*. Oxford University Press.
- Fernandez, Jordi. 2013. *Transparent Minds*. Oxford University Press.
- Finkelstein, David H. 2003. *Expression and the Inner*. Cambridge, MA: Harvard University Press. ISBN 9780674011564.
- Freidin, Robert. 2012. *Syntax: Basic Concepts and Applications*. Cambridge University Press.
- Gallois, André. 1996. *The World Without, the Mind Within: An Essay on First-Person Authority*. Cambridge University Press.
- Gazzaniga, Michael S. 1995. “Consciousness and the Cerebral Hemispheres.” In Michael S. Gazzaniga (ed.), *The Cognitive Neurosciences*. MIT Press.
- Gazzaniga, Michael S. and LeDoux, Joseph E. 1978. *The Integrated Mind*. Springer.
- Gertler, B. 2009. “Self-knowledge and the transparency of belief.” *Self-Knowledge. Oxford: Oxford* .
- Gertler, Brie. 2011. *Self-Knowledge*. Routledge.

- Gillon, Brendan S. 1990. "Ambiguity, Generality, and Indeterminacy: Tests and Definitions." *Synthese* 85:391–416.
- Goldman, Alvin. 1993. "The Psychology of Folk Psychology." *Behavioral and Brain Sciences* 16:15–28.
- . 1995. "Interpretation Psychologized." In Martin Davies and Tony Stone (eds.), *Folk Psychology*. Blackwell.
- . 2006. *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press, USA.
- Gordon, Robert M. 1995. "Folk Psychology as Simulation." In Martin Davies and Tony Stone (eds.), *Folk Psychology*. Blackwell.
- Grice, Paul. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Groenendijk, J and Stokhof, M. 1982. "The Semantic Analysis of WH-Compliments." *Linguistics and Philosophy* 5:175–233.
- Hamblin, C. L. 1958. "Questions." *Australasian Journal of Philosophy* 36:159–168.
- Hampshire, Stuart. 1975. *Freedom of the Individual*. Chatto & Windus.
- . 1982. *Thought and Action*. Chatto & Windus.
- Hampshire, Stuart and Hart, H. L. A. 1958. "Decision, Intention and Certainty." *Mind* LXVII.
- Heal, Jane. 1995. "Replication and Functionalism." In Martin Davies and Tony Stone (eds.), *Folk Psychology*. Blackwell.
- Hempel, Carl G. 1961. "Rational Action." *Proceedings and Addresses of the American Philosophical Association* 35:5–23.
- . 1965. *Aspects of Scientific Explanation And Other Essays in the Philosophy of Science*. The Free Press, New York.
- Higginbotham, J. 1993. "Interrogatives." In *The View From Building 20*. MIT Press.

- Hornsby, Jennifer. 2008. "A Disjunctive Conception of Acting for Reasons." In Adrian Haddock and Fiona Macpherson (eds.), *Disjunctivism: Perception, Action, Knowledge*. Oxford University Press.
- Howe, Reed B. K. 1991. "Introspection: A Reassessment." *New Ideas in Psychology* 9:25–44.
- Huddleston, Rodney and Pullum, Geoffrey K. 2005. *A Student's Introduction to English Grammar*. Cambridge University Press.
- Jackendoff, Ray. 1972. *Semantic Interpretation in Generative Grammar*. Studies in Linguistics. Cambridge, Massachusetts and London, England: The MIT Press.
- Johansson, Petter, Hall, Lars, Sikstrom, Sverker, and Olsson, Andreas. 2005. "Failure to Detect Mismatches Between Intention and Outcomes." *Science* 310:116–119.
- Johansson, Petter, Hall, Lars, Sikstrom, Sverker, Tarning, Betty, and Lind, Andreas. 2006. "How something can be said about telling more than we can know: on choice blindness and introspection." *Consciousness and Cognition* 15:673–692.
- Johansson, Petter, Hall, Lars, and Strandberg, Thomas. 2012. "Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey." *PLoS ONE* 7:1–8.
- Karttunen, Lauri. 1973. "Presuppositions of Compound Sentences." *Linguistic Inquiry* 4:169–193.
- Karttunen, L. 1977. "Syntax and Semantics of Questions." *Linguistics and Philosophy* 1:3–44.
- Kawamura, Tomoko. 2008. "Adverbial *Because*-clauses as Focal Elements." *Nanzan Linguistics* 2:103–122.
- Kraut, Robert E. and Lewis, Steven H. 1982. "Person Perception and Self-Awareness: Knowledge of Influences on One's Own Judgements." *Journal of Personality and Social Psychology* 42:448–460.

- Kripke, Saul. 2011. "Speaker's Reference and Semantic Reference." In *Philosophical Troubles*. Oxford University Press.
- Landman, Fred. 2000. *Events and Plurality: The Jerusalem Lectures*. Springer.
- Larson, Richard K. 2004. "Sentence-final Adverbs and "Scope"." *Proceedings of NELS 34* .
- Lawlor, Krista. 2009. "Knowing What One Wants." *Philosophy and Phenomenological Research* 79:47–75.
- Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.
- Lormand, Eric. 1996. "Inner Sense until Proven Guilty." .
- Lyons, William. 1986. *The Disappearance of Introspection*. The MIT Press.
- Matsuoka, Mikinari. 2013. "On the Notion of Subject for Subject-Oriented Adverbs." *Language* 89.
- McConnell-Ginet, Sally. 1982. "Adverbs and Logical Form: A Linguistically Realistic Theory." *Language* 58:144–184.
- Merchant, Jason. 2004. "Fragments and Ellipsis." *Linguistics and Philosophy* 27:661–738.
- Moran, Richard. 1988. "Making Up Your Mind: Self-Interpretation and Self-Constitution." *Ratio* 1:135–151.
- . 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton, New Jersey: Princeton University Press.
- . 2012. *Self-Knowledge, 'Transparency', and the Forms of Activity*. Oxford University Press.
- Morreal, John. 1979. "The Evidential Use of *Because*." *Papers in Linguistics* 12:231–238.

- Nichols, Shaun and Stich, Stephen P. 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. New York: Oxford University Press. ISBN 9780198236092.
- Nisbett, Richard E. and Bellows, Nancy. 1977. "Verbal Reports About Causal Influences on Social Judgements: Private Access Versus Public Theories." *Journal of Personality and Social Psychology* 35:613–624.
- Nisbett, Richard E. and Ross, Lee. 1980. *Human Inference: Strategies and Shortcomings of Social Judgement*. Prentice-Hall.
- Nisbett, Richard E. and Schachter, S. 1966. "Cognitive Manipulation of Pain." *Journal of Experimental Social Psychology* 2:227–236.
- Nisbett, Richard E. and Wilson, Timothy DeCamp. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84.
- Paul, Sarah K. 2009a. "How We Know What We're Doing." *Philosopher's Imprint* 9:1–24.
- . 2009b. "Intention, Beleif, and Wishful Thinking: Setiya on "Practical Knowledge"." *Ethics* 119.
- . 2012. "How we know what we intend." *Philosophical Studies* 161:327–346.
- . 2015. "The Transparency of Intention." *Philosophical Studies* 172:1529–1548.
- Peacocke, Christopher. 1999. *Being Known*. Oxford: Clarendon Press.
- Potts, Christopher. 2004. *The Logic of Conversational Implicatures*. Oxford University Press.
- . 2007. "Into the Conventional-Implicature Dimension." *Philosophy Compass* 2:665–679.
- Pryor, James. 2005. "There is Immediate Justification." In M Steup and E Sosa (eds.), *Contemporary Debates in Epistemology*. Oxford, Blackwell.
- Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.
- Reichenbach, Hans. 1971. *The Direction of Time*. University of California Press.

- Ryle, Gilbert. 1949. *The Concept of Mind*. Hutchinson.
- Sabini, John and Silver, Maury. 1981. "Introspection and Causal Accounts." *Journal of Personal and Social Psychology* 40:171–179.
- Sandis, Constantine. 2015. "Verbal Reports and 'Real' Reasons: Confabulation and Conflation." *Ethical Theory and Moral Practice* 18:267–280.
- Schnieder, Benjamin. 2010. "A Puzzle About 'Because'." *Logique et Analyse* 53:317–43.
- . 2011. "A Logic For 'Because'." *The Review of Symbolic Logic* 4:445–465.
- Schroeder, Mark. 2008. "Having Reasons." *Philosophical Studies* 139:57–71.
- . 2011a. "Holism, Weight, and Undercutting." *Noûs* 45:238–344.
- . 2011b. "What does it take to "have" a reason?" In A. Steglich-Petersen and A. Reisner (eds.), *Reasons for Belief*. Oxford University Press.
- . 2012. "Stakes, withholding, and pragmatic encroachment on knowledge." *Philosophical Studies* 160.
- . 2015. "In Defence of the Kantian Account of Knowledge: Reply to Whiting." *Logos & Episteme* .
- Schwitzgebel, Eric. 2014. "Introspection." *Stanford Encyclopedia of Philosophy* .
- Sennett, Adam. 2014. "Ambiguity." In *Stanford Encyclopedia of Philosophy*. Stanford University.
- Setiya, Kieran. 2003. "Explaining Action." *The Philosophical Review* 112:339–393.
- . 2007. *Reasons Without Rationalism*. Princeton University Press.
- . 2008. "Practical Knowledge." *Ethics* 118:388–409.
- . 2009a. "Practical Knowledge Revisited." *Ethics* 120:128–137.
- . 2009b. "Reasons and Causes." *European Journal of Philosophy* 19:129–157.

- . 2013. “Epistemic Agency: Some Doubts.” *Philosophical Issues* 23:179–198.
- Shah, Nishi and Velleman, J. David. 2005. “Doxastic Deliberation.” *Philosophical Review* 114:497–534.
- Shoemaker, Sydney. 1996. *The First-Person Perspective and Other Essays*. Cambridge, Cambridge University Press.
- . 2003. “Moran on Self-Knowledge.” *European Journal of Philosophy* 3:391–401.
- . 2012. “Self-Intimation and Second Order Belief.” In Declan Smithies and Daniel Stoljar (eds.), *Introspection and Consciousness*. Oxford University Press.
- Siegel, Susanna. 2010. *The Contents of Visual Experience*. Oxford University Press.
- Smith, Eliot R. and Miller, Frederick. 1978. “Limits on Perception of Cognitive Processes: a Reply to Nisbett and Wilson.” *Psychological Review* 85:355–362.
- Smith, Michael. 2004. “The Possibility of Philosophy of Action.” In *Ethics and the a priori: Selected essays on moral psychology and meta-ethics*. Cambridge University Press.
- Sosa, Ernest. 2003. “Does Knowledge Have Foundations?” In *Epistemic Justification: Internalism vs. Externalism, Foundations vs. Virtues*. Blackwell Publishing.
- Sprangers, Mirjam, van den Brink, Wulfert, van Heerden, Jaap, and Hoogstraten, Johan. 1987. “A Constructive Replication of White’s Alleged Refutation of Nisbett and Wilson and of Bem: Limitations on Verbal Reports of Internal Events.” *Journal of Experimental Social Psychology* 23:302–310.
- Stalnaker, Robert C. 1999a. “Pragmatic Presupposition.” In *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press.
- . 1999b. “Pragmatics.” In *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press.
- Stanley, Jason. 2011. *Know How*. Oxford University Press.

- Stanley, Jason and Williamson, Timothy. 2001. "Knowing How." *The Journal of Philosophy* 98:411–444.
- Stich, Stephen and Nichols, Shaun. 1995. "Folk Psychology: Simulation or Tacit Theory." In Martin Davies and Tony Stone (eds.), *Folk Psychology*. Blackwell.
- Stoljar, Daniel. 2012. "Introspective Knowledge of Negative Facts." *Philosophical Perspectives: Philosophy of Mind* 26:389–410.
- Storms, M. D. and Nisbett, Richard E. 1970. "Insomnia and the attribution process." *Journal of Personality and Social Psychology* 2:319–328.
- Thomason, Richmond H. and Stalnaker, Robert C. 1973. "A Semantic Theory of Adverbs." *Linguistic Inquiry* 4:195–220.
- van Fraassen, Bas. 1980. *The Scientific Image*. Oxford: Clarendon Press.
- Way, Jonathan. 2007. "Self-Knowledge and the Limits of Transparency." *Analysis* 16:223–230.
- Whately, Richard. 1855. *Elements of Logic*. James Munroe and Company.
- White, Peter A. 1980. "Limitations on Verbal Reports of Internal Events: A Refutation of Nisbett and Wilson and of Bem." *Psychological Review* 1:105–112.
- . 1987. "Causal Report Accuracy: Retrospect and Prospect." *Journal of Experimental Social Psychology* 23:311–315.
- . 1988. "Knowing more about what we can tell: 'Introspective access' and causal report accuracy 10 years later." *British Journal of Psychology* 79:13–45.
- Wilson, Timothy DeCamp. 2002. *Strangers to Ourselves*. Cambridge, Massachusetts: Belknap Press.
- Wilson, Timothy DeCamp and Nisbett, Richard E. 1978. "The Accuracy of Verbal Reports About the Effects of Stimuli on Evaluations and Behavior." *Social Psychology* 41:118–131.

- Wittgenstein, Ludwig. 1980. *Remarks on the Philosophy of Psychology*, volume 1. The University of Chicago Press.
- Wright, Crispin. 1998. "Self-knowledge: the Wittgensteinian legacy." *Knowing our own minds* 101–122.
- Wright, Crispin, Smith, Barry C.C., and MacDonald, Cynthia. (eds.). 1998. *Knowing our own Minds*. Oxford University Press, USA.
- Wright, Peter and Rip, Peter D. 1981. "Retrospective Reports on the Causes of Decisions." *Journal of Personal and Social Psychology* 40:601–614.
- Wyner, Adam Zachary. 1998. "Subject-Oriented Adverbs are Thematically Dependent." In Susan Rothstein (ed.), *Events and Grammar*. Springer.
- Zimmerman, Aaron Z. 2006. "Basic self-knowledge: Answering Peacocke's criticisms of constitutivism." *Philosophical studies* 128:337–379.
- . 2008. "Self-Knowledge: Rationalism vs. Empiricism." *Philosophy Compass* 3:325–352.
- Zwicky, A. M. and Sadock, J. M. 1975. "Ambiguity Tests and How to Fail Them." In Kimball (ed.), *Syntax and Semantics* 4. NY: Academic Press.