Knowing Why

Ryan Cox

ABSTRACT

In this essay, I argue that we have a non-inferential way of knowing particular explanations of our own actions and attitudes. I begin by explicating and evaluating Nisbett and Wilson's influential argument to the contrary. I argue that Nisbett and Wilson's claim that we arrive at such explanations of our own actions and attitudes by inference is not adequately supported by their findings because they overlook an important alternative explanation of those findings. I explicate and defend such an alternative explanation of how we can know such explanations in a non-inferential way, drawing on recent work in the philosophy of self-knowledge.

Please cite the published version of this paper. This version may contain typographical and typesetting errors. Published as Cox, Ryan. 2018. "Knowing Why." *Mind & Language* 33 (2): 177–97. https://doi.org/10.1007/s11229-019-02364-w

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 particular explanatory facts in a non-inferential way, just as we seem to know particular non-explanatory facts about our own actions and attitudes in a non-inferential way. I seem to know, in a non-inferential way, for instance, that I am going to the pub because Jane is there. And I seem to know, in a non-inferential way, that I believe that Jane is at the pub because her car is in the car park. If we know particular explanations of our own actions

and attitudes in a non-inferential way, then the problem of explaining how we know them 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, 18)[1]

The thought here is that in order for us 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—Davidson says we usually have *no* evidence or reasons—our beliefs cannot be *based* on such reasons, and, therefore, we do not *know* particular explanations of our own actions and attitudes by inference.

However, many theorists deny that we know such explanations in a non-inferential way. They are impressed by symmetries between the explanations we give of our own actions and attitudes and the explanations given by others. The symmetries, if they exist, are impressive. One particular symmetry, which is the focus of the famous studies of Nisbett and Wilson (1977), 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 very similar 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. 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. Many theorists have been persuaded by this simple argument and are happy to narrow the scope of the problem of self-

[1] This claim is endorsed by several contemporary theorists. James Pryor writes '...your beliefs about the reasons for which you acted are not ordinarily based on any evidence' (Pryor 2005, 533). Kieran Setiya writes: '...I know what my *reasons* are without having to find out' (Setiya 2007, 40). For similar claims see: (Setiya 2013, 192; Sandis 2015; Baker 2015, 3046).

knowledge to that of explaining how we know particular non-explanatory facts about our own actions and attitudes.[2]

So, on the one hand, we seem to have good reasons for thinking that we know particular explanations of our actions and attitudes in a non-inferential way *and*, on the other hand, we seem to have 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 and those who follow him are 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 inference hypothesis, as we might call it, over some alternative non-inferential hypothesis.

In this essay I defend the 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. In the final sections I develop such an alternative according to which our knowledge of particular explanations of our actions and attitudes is arrived at by resolving particular deliberative 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 actions and attitudes.

[2] This 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, 106). See also (Nisbett and Ross 1980, 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, 163). Discussing various empirical studies, Eric Schwitzgebel writes that the psychological literature '...can 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, sec. 4.2.1).

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. Strikingly, although Nisbett and Wilson's essay is often cited in support of the claim that we know particular explanations of our actions and attitudes by inference, 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 we make about particular explanations of our *own* actions and attitudes are just the same kind of errors someone else would make 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 others 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 draw on a similar argument made by Daryl Bem concerning our knowledge of our attitudes (Nisbett and Wilson 1977, 248; Bem 1967). Bem had argued that subjects showed the same kind of errors and ignorance about their own attitudes as suitably placed observers. He inferred from this that the subjects must be arriving at beliefs about their 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 a actor subject of attitudes to herself, and the ascriptions of an external observer with access to much the same information as actor subject, 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, 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, 186). Here are Nisbett and Wilson making a perfectly analogous 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, 247–48), quoting from (Bem 1967, 186)

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. In the remainder of this section, I explicate the introspection hypothesis and the inference hypothesis in detail, and argue that the inference hypothesis provides a better explanation of the symmetries than the inner-observation hypothesis.

2.1 The Inner-Observation Hypothesis

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, 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

[3] As 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, 17).

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, 232). And it is suggested again later in the essay when they answer this question:

> 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, 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.[4] Let's call the alternative hypothesis, then, the inner-observation hypothesis. 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, 246). Importantly, nothing about the inner-observation model predicts that introspection will be particularly reliable or error free.

It may seem striking that Nisbett and Wilson consider a kind of inner-observation 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, 233)

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

Goldman and others think that the implausibility of an inner-observation model of introspection as applied to knowledge of particular explanations of our own actions and attitudes can be established *a priori*. He writes:

[4] See (Lyons 1986) for a discussion of the history of inner-observation models of introspection. See (Shoemaker 1996) for some critical discussion.

> 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, 27)

If the inner-observation 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 (c.f. Beebee 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 (c.f. Anscombe 1963). Moreover, Nisbett and Wilson do not say that the inner-observation model of introspection is committed to the claim that we can know particular explanatory facts about our own actions and attitudes *directly* by inner observation. Rather, they seem to think that what we know by a kind of inner-observation is that some higher-order cognitive process has taken place and from this we arrive at beliefs about explanations. 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.

2.2 The Inference Hypothesis

The alternative to the introspection hypothesis, and the hypothesis that Nisbett and Wilson ultimately defend, is the inference hypothesis. According to the inference hypothesis, we know particular explanations of our actions and attitudes by inference. Nisbett and Wilson similarly do not offer a technical definition of inference, but we can assume that they intend their conclusion to be about the ordinary notion of inference.

According to the the ordinary notion of inference, 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—in a distinct but related sense—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 different conception of inference, one which departs either psychologically, or epistemologically, or in both ways, from the ordinary notion.

2.3 Why the Symmetries Support the Inference Hypothesis

Understood this way, it is clear that the inference hypothesis 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. Notice that the point here is not just that we are sometimes mistaken about the explanations of our own actions and attitudes. That wouldn't count against the inner-observation hypothesis. It is the kind of mistake which counts.

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 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.).

So it seems that *if* Nisbett and Wilson are right, and the symmetries exists, and the only potential explanations of the symmetries are the introspection hypothesis and the inference hypothesis, then the the conclusion that we know particular explanations of our actions and attitudes is well supported: it is the best explanation of the symmetries. For the remainder of the essay, I will evaluate this argument. In the next section I will look at the support for the major empirical premise of the argument. In the final sections of the essay 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 of the symmetries available.[5]

[5] I have attributed to Nisbett and Wilson a kind of inference to the best explanation. However, it might be thought that Nisbett and Wilson's argument is an argument by elimination. Perhaps Nisbett and Wilson are arguing that the only possible alternatives here are inference and introspection on the inner-observation model. It might be thought that their argument against the inner-observation 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-observation model and the inference model as the only alternatives. But they do so within the context of an argu-

3 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. 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, 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, 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

ment to the best explanation, not an argument by elimination. It is plausible to assume that the inner-observation 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-observation 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-observation model is correct. The overall structure of their argument remains an inference to the best explanation.

causal theories' (Nisbett and Bellows 1977, 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, 245).[6] 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, 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, 246).[7] 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.

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

[6] Also discussed in (Nisbett and Ross 1980, 209) and at greater length in (Wilson and Nisbett 1978, 124–26).

[7] Also discussed in (Nisbett and Ross 1980, 210) and at greater length in (Wilson and Nisbett 1978, 126–28).

an effect and of these, two thirds believed the phrase would have increased their willingness to take shocks' (Wilson and Nisbett 1978, 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, 249–50; Nisbett and Bellows 1977).[8]

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, 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 academic 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, 618). These subjects were not given any

[8] For further discussion see (Nisbett and Ross 1980, 212–16). See (Smith and Miller 1978, 358) for a critical discussion of some of the inferences involved in this study.

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, 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; and (iii) 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, 621–22)

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, 108–9). At this point there are two ways of responding to the argument. We can either question the evidence for the major empirical premise of the argument, or we can question the inference itself. For the remainder of this section I examine objections to the major empirical premise. For the remaining sections of the essay I examine the inference, arguing that it overlooks an alternative explanation of the symmetry.

3.1 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.[9] 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.[10]

The obvious problem with this kind of between-subject design is that it is compatible with the reassurance having no influence on the actor subjects. It is compatible with the fact, say, that 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

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

[10] 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)

> manipulations, particular subjects may have reported accurately the influences on their own judgements. (Nisbett and Ross 1980, 215)

Several critics of Nisbett and Wilson have drawn attention to this (Smith and Miller 1978; White 1980, 109–10; Wright and Rip 1981, 602).[11] 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, 357; White 1988, 21). Nobody 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.[12] Recall the job applicant study.

[11] Kraut 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, 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, 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, 359). See (Nisbett and Ross 1980, 215) for a response. See also (Howe 1991).

[12] This makes sense of their interest in the study by (Nisbett and Schachter 1966), reported in (Nisbett and Wilson 1977, 327), and in the study by (Storms and Nisbett 1970), reported in (Nisbett and Wilson 1977, 237–38).

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 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, 270).

I have some sympathy with this objection. But Sandis has cherrypicked his case. Perhaps 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

This also makes sense of Nisbett and Wilson's interest in subliminal perception and problem solving (Nisbett and Wilson 1977, 239–41).

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).[13] 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 Against The Inference Hypothesis

Before turning to an alternative explanation of the self and other symmetry, I want to examine the case for thinking we know explanations of our own actions and attitudes non-inferential in more detail. This will put more pressure on us to find an alternative explanation which reconciles the appearances, and the particular observation we will make will point us towards a particular alternative explanation of the appearances.

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.[14] This was Davidson's point. But there are further reasons for thinking that we do not arrive at such

[13] Thus we can agree with Victoria McGeer when she writes: '...what this research seems to be showing is precisely what tradition counter predicts, namely, that there are similar patterns of error across first- and third-person attributions. In other words, not only do we go wrong about ourselves; but when we go wrong, we tend to do so in exactly the same ways as we go wrong about others' (McGeer 1996, 491).

[14] Of course, this argument only works if we assume that in many case we *do* know particular explanatory facts about our actions and attitudes and that the means by which we know them is also operative in the cases of error that Nisbett and Wilson isolate. This seems like a reasonable assumption to make. (C.f. Nisbett and Ross: "This view of the origins of people's causal accounts does not apply merely to cases in which such accounts are inaccurate. It applies also to cases in which such accounts are accurate" (Nisbett and Ross 1980, 211)). Moreover, the argument assumes that in order to know something by inference you need sufficient reasons for believing it. Some theorists hold that you can know something by inference, even if you do not have sufficient reasons for so believing. So one could respond to the argument by defending this claim. I don't think either of these responses is particularly plausible, but do not have the space to argue for this conclusion here.

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, for that is how Nisbett and Wilson's conclusion is usually understood.

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, 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.

Moreover, when I know that I am going to the pub because Jane is there, I may not know that I am *aware* that Jane is there at all. Rather, my thoughts are directed entirely at the world, to facts like the fact that Jane is at the pub. It would be quite a leap to infer from the fact that I am going to the pub and the fact that Jane is at the pub that I am going to the pub because Jane is there. A crucial premise is missing, namely, the premise that I am aware that Jane is at the pub. And finally, we often find out *why* we are doing what we are doing at the same time that we find out *that* we are doing it. 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, vol. 1, sec. 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, 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 1982; Gallois 1996; Moran 2001; Byrne 2005; Fernandez 2013).[15] Of course, one might argue that sometimes, in order to know what I believe, I

[15] I should note that Byrne appeals to this observation in motivating an inferential account of self-knowledge. However, Byrne's inferential account appeals to a non-standard conception of inference and for our purposes can be thought of as a non-inferential account of self-knowledge.

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.

So there are good reasons for thinking that, typically, 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.

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.[16] 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.

5 An Alternative Explanation

As we have seen, much of the plausibility of Nisbett and Wilson's argument derives from the assumption that there are only two relevant alternative hypotheses concerning how we know particular explanations of our own actions and attitudes: the inner-observation hypothesis and the inferential hypothesis. In this section I want to introduce an alternative explanation of how we know particular explanations of our own actions and attitudes. In recent years, a kind of theory of self-knowledge has emerged as a rival to both inferential and inner-observation theories of self-knowledge. We can call this kind of

[16] Perhaps they are arriving at explanations by a kind of simulation rather than an inference. (See: (Heal 1995, 47; Gordon 1995, 71; Goldman 1995, 78). For critical discussion of the simulation theory see: (Stich and Nichols 1995, 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, 1995b).

theory an interrogative theory of self-knowledge, for reasons which will become apparent in a moment. In this section, I sketch such an interrogative theory of self-knowledge and offer some initial motivation for it. I will then argue that it offers an alternative explanation of the self and other symmetry.[17]

5.1 An Interrogative Theory of Self-Knowledge

Interrogative theories of self-knowledge hold that we can come to know the answers to particular theoretical questions about our own attitudes and actions by resolving distinct but related questions about our own actions and attitudes. For example, an interrogative theory of introspection might hold that you can come to know whether you will go to the party by resolving the question of whether to go to the party.

Several contemporary theories of self-knowledge can be thought of as interrogative theories of self-knowledge, although they are not thought of in such terms by their proponents.[18] All interrogative theories have the basic idea above in common, however they differ in their details. According to the interrogative theory which I think is most promising, we can come to know the answers to particular *theoretical* questions about our own attitudes and actions by resolving distinct but related *deliberative* questions about our own attitudes.[19] According to this theory, I can come to know whether I believe that Jane smokes—the answer to a theoretical question—by resolving the question of whether to believe that Jane smokes—a deliberative question. The relevant alternative interrogative theories hold that I can come to know whether I believe that Jane smokes by resolving the question of whether Jane smokes—a theoretical question—or by resolving the question of whether I should believe that Jane smokes—a normative question. On the face of it, these are distinct views, since the questions of whether to believe that Jane smokes, whether Jane smokes, and whether I should believe that Jane smokes, are distinct.

Henceforth I will simply speak of the interrogative theory of introspection, and mean by this a theory which holds that we can come to know the answer to particular theoretical questions about our own attitudes and actions by resolving distinct but related

[17] The 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 decision in non-inferential knowledge of action. Related ideas are developed in (Moran 1988, 2001, 2012; McGeer 1996, 2008).

[18] The theories of (Evans 1982) and (Moran 1988, 2001, 2012) are perhaps closest examples to pure interrogative theories of self-knowledge.

[19] The terms 'theoretical' and 'deliberative' are from (Moran 2001). As I use the terms they apply stipulatively to the intuitive distinction between questions like the question of whether *to* go to the party and questions like the question of whether one *will* go to the party.

deliberative questions about our own attitudes and actions.

According to the interrogative theory of introspection you can resolve a theoretical question about your own actions and attitudes merely by resolving a distinct but related deliberative question about your actions and attitudes. Resolving a question here should be thought of as a psychological process of some kind. Resolving a theoretical question may simply be a matter of coming to believe a particular answer to that question. Resolving some deliberative questions—for example, the deliberative question of whether to go to the party—may simply be a matter of forming an intention or making a decision. According to the interrogative theory you can resolve a theoretical question about your actions and attitudes merely by resolving a deliberative question about your actions and attitudes. This is what makes the interrogative theory a non-inferential theory. In order to resolve a theoretical question by resolving a deliberative question you needn't make an inference of any kind. Most importantly, you do not first resolve the deliberative question and then infer some answer to the relevant theoretical question from the fact that you have resolved the deliberative question.

Those are the basic elements of the interrogative theory of self-knowledge. The theory is perhaps most naturally applied in the case of knowledge of future action. You might come to know that you will go to the party by resolving the deliberative question of whether to go to the party. You might resolve the question of whether you will go to the party merely by resolving the question of whether to go to the party. This has seemed to many to be a natural theory of how we can come to know facts about our own future actions (c.f. Hampshire and Hart 1958; Hampshire 1975). Indeed, the interrogative theory is well-placed to explain the observations from the previous section, since it predicts that in coming to know whether you will go to the party, you will consider reasons for and against going to the party, not reasons for and against believing that you will go to the party. This is because resolving the question of whether to go to the party will often involve considering reasons for and against going to the party. And, moreover, not only will you not consider reasons for and against believing that you will go to the party, such reasons are entirely unnecessary. You can resolve the question of whether to go to the party without having good reasons for believing that you will go to the party.

Many theorists have thought that a similar theory can be offered in the case of knowledge of our attitudes (Evans 1982; Moran 2001, 2012). You might come to know that you believe that Jane is at the party by resolving the question of whether to believe that Jane is at the party. You can resolve the *theoretical* question of whether you believe that Jane is at

the party by resolving the *deliberative* question of whether to believe that Jane is at the party. And you might resolve the deliberative question of whether to believe that Jane is at the party, in turn, by simply resolving the theoretical question of whether Jane is at the party.[20] The interrogative theory is well placed to explain the observations from the previous section, since it predicts that in coming to know whether you believe that Jane is at the party, you will consider reasons for and against believing that Jane is at the party, and not reasons for and against believing that that you believe that Jane is at the party. Moreover, not only will you not consider reasons for and against believing that you believe that Jane is at the party, such reasons are entirely unnecessary. You can resolve the question of whether to believe that Jane is at the party without having good reasons for believing that you believe that Jane is at the party.

A natural question arises at this point as to whether you can come to *know* that you will go to the party or come to know that you believe that Jane is at the party in this way. But since there is a strong presumption that we typically do know these things, then any theory of how we are arriving at the relevant beliefs is a theory of how we can come to know these things.[21] To challenge the presumption that we know these things by arguing that we can only know such things by inference from good reasons is to beg the question against any non-inferential theory.

Now I think that one serious motivation for the interrogative theory (at least the version that we are considering here) is the ease with which it can be extended to give an account of how we know particular explanatory facts about our own actions and attitudes. According to the interrogative theory, we can come to know particular explanatory facts about our own actions and attitudes by resolving distinct but related deliberative questions about our actions and attitudes. The question we face is: which deliberative questions? I want to suggest that we can come to know particular explanations of our own actions and attitude by resolving the deliberative question of whether to treat some fact

[20] Resolving the question of whether to believe that Jane is at the party isn't equivalent to resolving the question of whether Jane is at the party. You might resolve the question of whether to believe that Jane is at the party without resolving the question of whether Jane is at the party, for you might resolve to withhold belief concerning whether Jane is at the party.

[21] One might wonder whether Nisbett and Wilson would accept this presumption. Surprisingly, there is some evidence that they would. Consider the following passage from Nisbett and Ross: "People are, despite Nisbett's and Wilson's demonstrations, often right in their accounts of the reasons for their behavior. A person who answers a telephone and asserts that he did so"because it was ringing" is surely right. A person who solves a problem by applying an appropriate algorithm and then asserts that he solved the problem by applying the algorithm, is right. A person who asserts that he opened the refrigerator door because he was hungry is usually right. But we have theories about why we answer telephones, how we solve problems, and why we open refrigerators, and these theories are usually correct. Because these theories are so manifestly correct, however, it would require some ingenuity to find cases in which the knowledgeable observer was not also correct" (Nisbett and Ross 1980, 211).

(or, more generally, some consideration) as a reason for performing that action or having that attitude. For example, according to this theory, I can come to know whether I will go to the party because Jane is there by resolving the deliberative question of whether to treat the fact that Jane is at the party as a reason to go to the party. The fact that Jane is at the party is a reason for me to go to the party, but it is a further matter as to whether I treat it as a reason to go to the party.

That is the suggestion. It is a natural extension of the interrogative theory of self-knowledge as I have developed it here. Indeed, the interrogative theory seems to require such an extension. Resolving the question of whether to go to the party, for example, involves not only considering reasons for and against going to the party, but resolving the question of whether to treat those reasons as reasons for or against going to the party. The fact that Jane is at the party might be a reason for me to go to the party, and the fact that John is at the party might be a reason for me to go to the party. I might consider both of these reasons in resolving the question of whether to go to the party, but I may resolve the question of whether to go to the party for one of these reasons and not for the other. What makes the difference is whether I treat the fact that Jane is at the party or whether I treat the fact that John is at the party as a reason for going to the party.

The interrogative theory is, again, well placed to explain the kind of observations we made in the previous section, since it predicts that when I come to know whether I am going to the party because Jane is there, I consider reasons for and against going to the party, reasons like the fact that Jane is there, and the fact that John is there, and I do not consider reasons for believing that I am going to the party because Jane is there. Moreover, according to the interrogative theory, such reasons are irrelevant. In order to resolve the question of whether to treat some fact (or consideration) as a reason for performing some action, I do not need good reasons for believing that I will perform that action for that reason.

Let me emphasise at this point that resolving the deliberative question of whether to treat some consideration as a reason is not equivalent to the theoretical question of whether some consideration is a reason, and nor is it equivalent to the normative (theoretical) question of whether I should treat that consideration as a reason.[22] If it was, then resolving the question of whether to treat some consideration as a reason would involve considering reasons for and against believing that it is a reason, and would end either in a belief that it is, or a belief that it isn't, or it would involve considering reasons for and

[22] This sets my view apart from the account in (Setiya 2013).

against believing that I should treat that consideration as a reason, and would end either in a belief that I should, or a belief that I shouldn't. In contrast, if resolving the question of whether to treat some consideration as a reason involves considering further reasons at all, it involves considering reasons for and against treating the consideration as a reason.[23]

5.2 Explaining the Symmetry

We are now in a position to offer an alternative explanation of Nisbett and Wilson's symmetry. The key to the explanation is the following observation. The actor subject resolves the question of whether some consideration is the reason for which they made the choice or evaluation they did by resolving the question of whether to *treat* that reason as a reason for making the choice or making the evaluation in question. In contrast, the observer resolves the question of whether some consideration is the reason for which the actor subject make her choice or evaluation by resolving the question of whether the subject resolved the question of whether to treat the consideration as a reason for making the choice or making the evaluation in question in the affirmative or in the negative.[24] The observer assumes that the subject is rational to a certain degree and so assumes that the subject's choices and evaluations are explained by the reasons the subject has resolved to treat as reasons. The actor subject need not make any such assumption about herself. Instead, she simply resolves the question of whether to treat some consideration as a reason for making some choice or making some evaluation, and resolves the question of whether that reason *is* her reason, by resolving the question of whether to treat it as her reason.

But then why should two distinct processes lead to the same explanations, as symmetry predicts? Well, this is because the observer's reasoning is guided by the assumption that the subject is rational. The observer assumes that if something is a good reason for some response, then, at least typically, the subject will treat it as a reason for her response, and that if something is a bad reason for some response, then, at least typically, the subject will not treat it as a reason for her response. Now, the subject's reasoning is not, itself,

[23] This psychological picture is particularly well suited to views of reasons according to which there are reasons for and against treating other reasons as reasons. There might seem to be a problematic regress here. But there isn't. At some point, one just treats some consideration as a reason without resolving the question of whether some further consideration is a reason for treating it as a reason.

[24] We would need an alternative explanation if we assume that observers are arriving at their explanations by a kind of simulation. We might then argue that the process of simulation involves resolving questions about what to do and what to think in certain hypothetical situations—i.e. the situations of the actor subjects. If so, this would explain the symmetries.

guided by the assumption of rationality. Rather, the subject's reasoning simply *conforms* to the assumptions of rationality. If something is a good reason for some response, then, typically, the subject will treat it as a good reason for the response. If it is a bad reason for some response, then, typically, the subject will not treat it as a reason for that response. The upshot is that, although the subject and the observer use different means, they arrive at roughly the same explanations, just as symmetry predicts. Moreover, we get the more precise prediction that both the subject and the observer will tend to assume that the subject's responses are explained by things which are good reasons for those responses, even when this is not the correct explanation of the responses, and both subjects and observers will tend to deny that some consideration is an explanation of the response when that consideration isn't an obviously good reason for that response.

Let's see how the explanation works in more detail by considering an example. Consider Nisbett and Wilson's reassurance and willingness study again. Recall that whether a subject received a reassurance made no difference to how far they went in administering shocks to themselves. However, actor subjects reported that they went as far as they did because they received the reassurance. And observer subjects arrived at similar explanations. According to the interrogative theory, the actor subject is resolving the question of whether she went as far as she did because she received the reassurance by resolving the question of whether to treat the fact that she received the reassurance as a reason for administering more shocks. Since the fact that she received such a reassurance is a reason for going further than she otherwise might have, it isn't surprising that the actor subject arrives at the explanation that she did indeed go as far as she did because she received the reassurance. The observer subject reasonably assumes that the subject would have treated the fact that she received the reassurance as a reason for going as far as she did, and infers that it is the reason for which the subject did go as far as she did. Although the actor subject arrives at her explanation by a non-inferential means—by resolving a deliberative question—and the observer subject arrives at her explanation by inference, they both arrive at the same explanation, in a predicable way. And, if Nisbett and Wilson are right that this isn't the reason why the subject went as far as she did, then both the actor subject and the observer subject are in error.

The opposite kind of error can be illustrated by considering the job candidate study. Recall that in that study, actor subjects and observer subjects denied that the actor subjects judged that the job candidate was likeable because she was attractive. Yet, Nisbett and Wilson argue that the actor subject's judgements were influenced by the fact that the

candidate was attractive. This seems like a clear case where a rational explanation obtains, and yet both actor subjects and observer subject deny that it does. We can explain how the actor subject arrives at her explanation on the hypothesis that she resolves the question of whether she judged that the candidate was likeable by resolving the question of whether to treat the fact that the candidate is attractive as a reason for judging that she is likeable. Since this fact is not a good reason for judging that the candidate is likeable, it isn't surprising that the actor subject denies that it is her reason for judging that the candidate is likeable. Similarly, the observer subject will assume that the subject will not treat this fact as a reason for judging that the candidate is likeable. Again, although the actor subject and the observer subject are arriving at their explanations by different means, the nature of those means ensures that they will arrive at roughly the same explanations.

We are now in a position to see that the interrogative 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 interrogative theory is the *best* explanation of these facts about our knowledge of such explanations of our own actions and attitudes. More needs to be done in explicating the interrogative 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 explanations of our actions and attitudes.

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.[25] Many theorists happily concede our knowledge of rational explanations to the inferential theory, while holding that doing so is not at odds with the claim that we each have a non-inferential way of knowing our own

[25] The 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, 2009) as well as its opponents (Paul 2009b, 2009a)

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, 75)[26]

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.

Whether or not the interrogative theory of self-knowledge I have sketched here turns out to be correct, we certainly have good reasons to look for alternative theories of self-knowledge which explain how we can know particular explanations of our own actions and attitudes in a non-inferential way, while at the same time explaining why our explanations pattern with the explanations of suitable placed observers.[27]

Bibliography

Alvarez, Maria. 2010. *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford University Press.

Anscombe, Elizabeth. 1963. *Intention*. Harvard University Press.

Baker, Derek. 2015. "Why Transparency Undermines Economy." *Synthese* 192 (9).

Beebee, Helen. 2003. "Seeing Causing." *Proceedings of Aristotelian Society* 103 (1).

Bem, Daryl J. 1967. "Self-Perception: An Alternative Interpretation of Cognitive Dissonance Phenomena." *Psychological Review* 74 (3): 183–200.

[26] She 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, 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.

[27] I'd like to thank David Chalmers, Luara Ferracioli, Daniel Gregory, Frank Jackson, Naomi Kloosterboer, Eric Llamas, Luke Roelofs, and Daniel Stoljar, for comments and discussion. I'd also like to thank audiences at the Current Projects Seminar at the University of Sydney, the PhilSoc Seminar at the Australian National University in 2015, and the 2014 Australasian Association of Philosophy Conference in Canberra, for comments on earlier versions of this material. Finally, I'd like to thank two anonymous referees for Mind & Language for their very helpful comments and suggestions.

Byrne, Alex. 2005. "Introspection." *Philosophical Topics* 33 (1): 79–104.

Davidson, Donald. 1980. *Essays on Actions and Events*. Oxford: Oxford University Press.

Davies, Martin, and Tony Stone, eds. 1995a. *Folk Psychology*. Blackwell.

———, eds. 1995b. *Mental Simulation*. Blackwell.

Ericsson, K. Anders, and Herbert A. Simon. 1980. "Verbal Reports as Data." *Psychological Review* 87 (3).

Evans, Gareth. 1982. *The Varieties of Reference*. Edited by John McDowell. Oxford University Press.

Fernandez, Jordi. 2013. *Transparent Minds*. Oxford 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 *The Cognitive Neurosciences*, edited by Michael S. Gazzaniga. MIT Press.

Gazzaniga, Michael S., and Joseph E. LeDoux. 1978. *The Integrated Mind*. Springer.

Gertler, Brie. 2011. *Self-Knowledge*. Routledge.

Goldman, Alvin. 1993. "The Psychology of Folk Psychology." *Behavioral and Brain Sciences* 16: 15–28.

———. 1995. "Interpretation Psychologized." In *Folk Psychology*, edited by Martin Davies and Tony Stone. Blackwell.

———. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press, USA.

Gordon, Robert M. 1995. "Folk Psychology as Simulation." In *Folk Psychology*, edited by Martin Davies and Tony Stone. Blackwell.

Hampshire, Stuart. 1975. *Freedom of the Individual*. Chatto & Windus.

———. 1982. *Thought and Action*. Chatto & Windus.

Hampshire, Stuart, and H. L. A. Hart. 1958. "Decision, Intention and Certainty." *Mind* LXVII (265).

Heal, Jane. 1995. "Replication and Functionalism." In *Folk Psychology*, edited by Martin Davies and Tony Stone. Blackwell.

Howe, Reed B. K. 1991. “Introspection: A Reassessment.” *New Ideas in Psychology* 9 (1): 25–44.

Johansson, Petter, Lars Hall, Sverker Sikstrom, and Andreas Olsson. 2005. “Failure to Detect Mismatches Between Intention and Outcomes.” *Science* 310: 116–19.

Johansson, Petter, Lars Hall, Sverker Sikstrom, Betty Tarning, and Andreas Lind. 2006. “How Something Can Be Said about Telling More Than We Can Know: On Choice Blindness and Introspection.” *Consciousness and Congnition* 15: 673–92.

Kraut, Robert E., and Steven H. Lewis. 1982. “Person Perception and Self-Awareness: Knowledge of Influences on One’s Own Judgements.” *Journal of Personality and Social Psychology* 42 (3): 448–60.

Lyons, William. 1986. *The Disappearance of Introspection*. The MIT Press.

McGeer, Victoria. 1996. “Is "Self-Knowledge" an Empirical Problem? Renegotiating the Space of Philosophical Explanation.” *The Journal of Philosophy* 93 (10): 483–515.

———. 2008. “The Moral Development of First-Person Authority.” *European Journal of Philosophy* 16 (1): 81–108.

Moran, Richard. 1988. “Making up Your Mind: Self-interpretation and Self-Constitution.” *Ratio* 1 (2): 135–51.

———. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton, New Jersey: Princeton Univeristy Press.

———. 2012. “Self-Knowledge, ‘Transparency,’ and the Forms of Activity.” In, edited by Daniel Stoljar and Declan Smithies. Oxford University Press.

Nichols, Shaun, and Stephen P. Stich. 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. New York: Oxford University Press.

Nisbett, Richard E., and Nancy Bellows. 1977. “Verbal Reports about Causal Influences on Social Judgements: Private Access Versus Public Theories.” *Journal of Personality and Social Psychology* 35 (9): 613–24.

Nisbett, Richard E., and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgement*. Prentice-Hall.

Nisbett, Richard E., and S Schachter. 1966. “Cognitive Manipulation of Pain.” *Journal of Experimental Social Psychology* 2: 227–36.

Nisbett, Richard E., and Timothy DeCamp Wilson. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84 (3).

Paul, Sarah K. 2009a. "How We Know What We're Doing." *Philosopher's Imprint* 9 (4): 1–24.

———. 2009b. "Intention, Belief, and Wishful Thinking: Setiya on 'Practical Knowledge'." *Ethics* 119.

Pryor, James. 2005. "There Is Immediate Justification." In *Contemporary Debates in Epistemology*, edited by M Steup and E Sosa. Oxford, Blackwell.

Sabini, John, and Maury Silver. 1981. "Introspection and Causal Accounts." *Journal of Personal and Social Psychology* 40 (1): 171–79.

Sandis, Constantine. 2015. "Verbal Reports and 'Real' Reasons: Confabulation and Conflation." *Ethical Theory and Moral Practice* 18: 267–80.

Schwitzgebel, Eric. 2014. "Introspection." *Stanford Encyclopedia of Philosophy*.

Setiya, Kieran. 2007. *Reasons Without Rationalism*. Princeton University Press.

———. 2008. "Practical Knowledge." *Ethics* 118: 388–409.

———. 2009. "Practical Knowledge Revisited." *Ethics* 120 (1): 128–37.

———. 2013. "Epistemic Agency: Some Doubts." *Philosophical Issues* 23: 179–98.

Shoemaker, Sydney. 1996. *The First-Person Perspective and Other Essays*. Cambridge, Cambridge University Press.

Siegel, Susanna. 2010. *The Contents of Visual Experience*. Oxford University Press.

Smith, Eliot R., and Frederick Miller. 1978. "Limits on Perception of Cognitive Processes: A Reply to Nisbett and Wilson." *Psychological Review* 85 (4): 355–62.

Sprangers, Mirjam, Wulfert van den Brink, Jaap van Heerden, and Johan Hoogstraten. 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: 302–10.

Stich, Stephen, and Shaun Nichols. 1995. "Folk Psychology: Simulation or Tacit Theory." In *Folk Psychology*, edited by Martin Davies and Tony Stone. Blackwelll.

Storms, M. D., and Richard E. Nisbett. 1970. "Insomnia and the Attribution Process." *Journal of Personality and Social Psychology* 2: 319–28.

White, Peter A. 1980. "Limitations on Verbal Reports of Internal Events: A Refutation of Nisbett and Wilson and of Bem." *Psychological Review* 1: 105–12.

———. 1987. "Causal Report Accuracy: Retrospect and Prospect." *Journal of Experimental Social Psychology* 23: 311–15.

———. 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 Richard E. Nisbett. 1978. "The Accuracy of Verbal Reports about the Effects of Simuli on Evaluations and Behavior." *Social Psychology* 41 (2): 118–31.

Wittgenstein, Ludwig. 1980. *Remarks on the Philosophy of Psychology*. Edited by G. E. M. Anscombe and G. H. von Wright. Vol. 1. The University of Chicago Press.

Wright, Peter, and Peter D. Rip. 1981. "Retrospective Reports on the Causes of Decisions." *Journal of Personal and Social Psychology* 40 (4): 601–14.