Judgment as a Guide to Belief[*]

Nicholas Silins
Cornell University

Introduction

What is the role of consciousness in our introspective lives? In this paper, I will focus on the role of conscious judgment in giving us access to our standing non-conscious beliefs. The view I will defend is a special case of a more general position in the epistemology of introspection. I will start by sketching the general view, and then give the details of my particular position.

One plank of the general position concerns specialness: our self-ascriptions of mental states can be justified in a way that our ascriptions of mental states to others are not.

Another plank of the position concerns fallibility: we can nevertheless make justified mistakes about what mental states we are in. For example, if you hallucinate that something yellow is present, with no indication that anything is going wrong, you can be justified in believing that you are in the relational mental state of *seeing* something yellow, even though you are not in that state.

The **moderate position** says that our introspective beliefs are sometimes justified in a both fallible and non-inferential way.[1] Accepting this position is, I think, an attractive way of

[*] Thanks to seminar participants at Cornell, and audiences at an Arché Basic Knowledge Workshop, Princeton, Columbia, the Hong Kong Towards a Science of Consciousness conference of 2009, and at the Research School of Social Sciences at the Australian National University. Thanks in particular to Paul Benacereff, Selim Berker, Alex Byrne, David Chalmers, John Collins, Shamik Dasgupta, Dylan Dodd, Alan Hájek, Gilbert Harman, Mark Johnston, Thomas Kelly, Brian Kim, Boris Kment, Maria Laasonen, Sarah-Jane Leslie, Errol Lord, Angela Mendelovici, Ram Neta, James Pryor, Carol Rovane, Danielle Sagravatti, Kranti Saran, Jonathan Schaffer, Kieran Setiya, Sydney Shoemaker, Ralph Wedgwood, Timothy Williamson, and Elia Zardini. Finally, thanks especially to Declan Smithies and Daniel Stoljar.

[1] In maintaining the moderate view, one need not say that we never have inferential access to our mental states, or that we never have infallible access to our mental states. For example, perhaps your pain can give you infallible justification to believe that you are in pain. Or

avoiding two bad views. According to one, we cannot be mistaken about our own minds; according to the other, there is nothing distinctive about our introspective beliefs.

The moderate position is attractive, but the debate in the area is often set up so as to leave it nearly out of sight. It is standard to ask about how we have self-*knowledge* or about how we have *access* to our mental states. Both these questions foreground the cases in which we are in a given mental state. But we need to consider what our positive epistemic position can be for a mistake about our own mind.

One might expect Descartes to be the archenemy of the moderate position, but he arguably is not. He commits himself to something close to it in the following passage, as it is translated by John Cottingham (the passage was drawn to my attention by Moran 2001: 12, n.9):

> many are themselves ignorant of their beliefs. For since the action of thought by means of which we believe something is different from that by means of which we know that we believe it, the one is often found without the other (*A Discourse on Method*, part 3).

I should start by clearing away a deflationary reading of the passage. On this reading, Descartes is merely saying that people sometimes have a belief without knowing they have it, say because they have not considered the question of whether they have the belief. This reading interprets the final phrase in the Cottingham translation as saying that "the former is often found without the latter". A look at the French resolves the problem. Descartes' own phrase is "elles sont souvent l'une sans l'autre", which should be rendered as "each one is often found without the other". So the passage in fact does make a quite striking claim: the action of thought by means of which we know that we have a belief (when all goes well) is often found without the action of thought by means of which we believe.

There are open questions about what Descartes means here by "action of thought". I will set them aside, and use Descartes' remark as a springboard to express the position I will defend in this paper (the position is developed in a related but different way by Peacocke 1999). The view is an instance of the moderate position.

---

perhaps you can have inferential justification to believe that you are in pain by inference from observation of your behavior (in addition to having justification to believe that you are in pain stemming from your pain itself).

I would first adapt what Descartes says as follows: the "action of thought" by which we know we believe something is sometimes our conscious judgment, where we can consciously judge that p without believing that p, and vice versa. The action of thought by which we sometimes know we believe that p thus does not guarantee the presence of a belief that p. Moreover, one can believe that p without judging that p---each state can be found without the other. Rather than focusing on knowledge, however, I prefer to focus on justification. That's because I am interested in what your positive epistemic position can be for a mistake about your own mind, and almost everyone agrees you can't know a false proposition about anything.[2] Finally, I would add in particular that conscious judgments give us non-inferential justification for second-order beliefs. Here I go well beyond what Descartes says.

My central thesis about conscious judgments is thus the following instance of the moderate position:

> (**Conscious Judgment**): Conscious judgments give us non-inferential yet fallible justification for second-order beliefs.

I will defend the thesis by looking closely at the famous "transparency method" discussed by Edgley (1969), Evans (1982), Moran (2001), and others. Very roughly, the key idea is that, when you answer the question whether p, you put yourself in a position to answer the question whether you believe that p. As I will develop the idea, judgment is a guide to belief. In particular, one's conscious judgments are a basic yet fallible guide to one's non-conscious standing beliefs.

There are several reasons why it is important to take a much closer look at the transparency method. Of course we need to understand the method to understand the epistemology of introspection more generally. But the topic is important for metaphysical reasons as well as epistemological reasons. By considering the role of conscious states as a guide to our non-conscious states, we should gain a valuable constraint on views of consciousness. Our understanding of what consciousness is must allow it to play the epistemic role it in fact plays. In the conclusion I will highlight two specific upshots of our discussion for the nature of consciousness.

---

[2] For dissent, see Hazlett (2010).

I will discuss the transparency method in the first half of this paper. Here I will argue that, on the best understanding of the transparency method, judgments give us non-inferential yet fallible justification for second-order beliefs. The second half will respond to important objections to the view. The main challenge comes from so-called "constitutivist" views in the epistemology of introspection, which instead emphasize the role of beliefs themselves in giving one justification for second-order beliefs. Here I will give a systematic survey of quite different ways of developing this approach. Only some constitutivist claims will turn out to be incompatible with my view. I will argue that each of those claims is false or unmotivated.

Before I take on the main projects of the paper, let me introduce some key terms and clarifications.

First of all, when I speak of "beliefs" in what follows, I will only have what one might call "standing beliefs" or "dispositional beliefs" in mind. You most likely had a standing belief a moment ago that the author of this paper is Silins, even though you weren't judging that the author of this paper is Silins.

When I speak of "judgments" in what follows, I will only have conscious judgments in mind, those which modify what it is like for you at the time you make them. I won't be concerned with non-conscious judgments (if there are any at all).[3] To make a judgment of the kind I am interested in, you might sincerely assert to someone that p. But you can consciously judge that p without performing the linguistic act of assertion. In many cases conscious judgment will be the "inner analogue" of assertion, although it may well be that one can judge that p without in any way vocalizing or imagining a sentence with the content that p.

On one view, judgments and beliefs are very closely related: a conscious judgment is simply a standing belief that has *become* conscious. One reason to doubt this view is that judgments are often caused by beliefs. For there to be causal relations between beliefs and judgments, we need two non-identical states which are causally related. In what follows we will also see a stronger form of distinctness between judgment and belief---important cases in

[3] One *might* count the following sort of case as one involving non-conscious judgment: you have an occurrent rather than dispositional belief, given the active role of the belief in guiding your action, yet the belief is not conscious, given that the belief does not by itself modify what it is like for you at the time.
I set aside the case of occurrent yet non-conscious belief in what follows.

which one judges that p yet does not believe that p. To anticipate, one can judge that p as a result of a slip which fails to reflect one's standing beliefs.

Second, let me clarify what I have in mind by "direct" or "non-inferential" access to one's mental states. Let's say that you have **immediate** justification to believe that p just in case you have justification to believe that p, and you do so in a way which does not rely on your justification to hold any other belief (Pryor 2005).

Notice that immediate justification is characterized in terms of *how* one gets to have it, rather than in terms of how strong it is, or in terms of when one gets to have it. The key requirement is that one is not made to have immediate justification by one's having justification for any further belief.

Third, let me clarify what I have in mind by "introspective justification": you have **introspective** justification to believe that you are in a mental state M just in case you have justification to believe that you are in M, and you do so in a way such that no one else can have justification to believe that you are in M in that way.[4]

Notice that introspective justification is characterized in terms of *who* it is available to, rather than in terms of how strong it is, or in terms of when exactly one gets to have it. I characterize introspective justification by its "peculiarity" (Byrne 2005), not by its superiority. I also do not characterize introspective justification with any positive account of how it is acquired. The characterization leaves open whether introspective justification is immediate or not, whether it is acquired by "inner sense" or not, and so on.

Finally, let me clarify what I have in mind by "fallible" access. I will say that a state j gives you **fallible** justification to believe that p just in case j gives you justification to believe that p and it is possible for you to be in j while it is not the case that p.[5]

---

[4] At a minimum, no one else actually has the ability to access your mental states in the relevant way. I leave open whether there is any stronger sense in which it is impossible for others to access your mental states in the relevant way.

[5] According to Sutton (2007), it is not possible to have a justified false belief. This extreme position is actually compatible with the claim that one sometimes has fallible justification. Our definition of "fallible" justification is actually silent about whether, when you have a fallible justification to believe that p, you could have justification to believe that p if it is not the case that p. The crucial question for our definition is whether a justifying state is such that one can be in it when it is not the case that p, leaving open whether it still gives justification to believe that p when it is not the case that p.

1. Transparency and Belief

A useful place to start in characterizing the transparency method is its famous discussion by Gareth Evans. We can learn a lot by reflecting on this discussion. However, much of what we will learn concerns how Evans is misleading or mistaken. He writes that

> in making a self-ascription of a belief, one's eyes are, so to speak, or occasionally literally, directed outward---upon the world. If someone ask me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p . . . If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states: even the most determined sceptic cannot find here a gap in which to insert his knife (Evans 1982, 225).

We can hint at the key lesson from this passage in the following way:

> (**Slogan**): You can answer the question whether you believe that p by answering the question whether p.

In order to make further progress, we need to consider several further questions. What is the method? When does it work? What does it do when it works? How does it work?

To gain a grip on what the method is, let's start by looking at Evans' claim that, "I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?'". This remark suggests that, in order to find out whether one believes that p, one has to launch a new enquiry into whether p. The suggestion has at least two problems.[6]

First of all, if you judge that p, without having launched a new investigation into the matter, you are still in a good position to answer the question whether you believe that p. You need not do what Evans seems to demand in order to answer the question about your mind.

Second, if you want to find out whether you already believed that p, you arguably *should* not do what Evans seems to demand, since a new enquiry might easily result in a new belief that p. For example, if you wondered whether you had a prior belief that God exists,

[6] For further discussion of related complications, see Peacocke (1998: 215-6), Shah and Velleman (2003: 506-8), and Byrne (2005: 84-5).

you shouldn't answer that question by considering considerations for and against the existence of God.

In what follows I will largely bracket questions about your access to your antecedent beliefs. My main focus will instead be on the situation once you do answer the question whether p, and how you stand with respect to your current beliefs once you do. Since I am bracketing questions about one's prior beliefs, I am also interested in the introspective upshot of answering the question whether p, whether or not you started out wondering about your beliefs before you asked the question whether p. We will therefore be concerned with a much wider range of cases than Evans.

Let me now turn to a different question---when you answer the question whether p, *which* epistemic position are you in with respect to whether you believe that p? Evans boldly asserts that his procedure is "necessarily" a source of knowledge. We have some reason to doubt the bold claim. One complication arises if someone believes each of two contradictory propositions (I assume it is possible for a person to do so). Here the person's belief that not-p might prompt her to answer the question whether p with a "no", and to answer the question whether she believes that p with a "no". Since she still believes the contradictory proposition that p, she does not achieve knowledge that she does not believe that p (we will see further problem cases in what follows).

A more cautious view is answering the question whether p necessarily gives one *justification* for a second-order belief. The proposal is more plausible than the knowledge proposal given that the current proposal is less demanding. Also, in the case of contradictory belief, the person arguably does gain justification to believe she doesn't believe that p.

There are two more wrinkles about what position you are in when you answer the question whether p. Notice that you can have justification to believe that p whether or not you in fact believe that p on the basis of the justification, and whether or not you in fact believe that p at all. On the view I will develop, answering the question whether p gives you justification whether or not you have a belief on the basis of that justification. In current jargon, I take answering the question whether p to provide what is known as "propositional justification" rather than "doxastic justification".

Let's now turn to the question of *when* answering the question whether p gives one justification. According to Edgley 1969 or Moran 2001, the question whether one believes that p is "transparent" to the question whether p. On this line of thought, presumably shared

by Evans himself, whatever the answer one gives to the question whether p, one will be in a good epistemic position to give the very same answer to the question whether one believes that p. This view exaggerates the success of the transparency method (whose name might indeed be a misnomer). Sometimes one's answer to the question whether p is just "maybe", for example if I ask myself whether it will rain one month from now. In many of those cases, however, one is still in a good position to answer the question of whether one believes that p with a "no", rather than with a "maybe". Strictly speaking, the question whether one believes that p is not transparent to the question whether p.

I will work with a more cautious claim about when answering the question whether p gives one justification---it gives one justification when one judges that p. I thus take Evans' procedure to be a guide to the presence of beliefs rather than to the absence of beliefs.[7] I would therefore describe the method as follows:

(**Transparency Method**): If you judge that p, believe you believe that p!

Given that my main interest is in your epistemic position when you judge that p, whether or not you have taken advantage of that position, I will focus on the following kind of formulation:

> (**Transparency Thesis, First Pass**): If you judge that p, then you have justification to believe that you believe that p.[8]

Although I have not formulated the thesis explicitly in terms of *prima facie* justification, please do read it and its successors with that in mind. When one judges that p, one's

[7] I don't take the method to be our *only* guide to our beliefs. Evans seems to think otherwise, given his remarks about what one "must" do to answer the question whether one believes that p. To see why there are arguably other sources of introspective justification, consider that you could have introspective justification to believe that p when you may never have judged that p. Perhaps your standing belief that p could give you introspective justification for a second-order belief, without doing so via an intermediary judgment (Zimmerman 2006).

[8] Although I won't discuss the further claims below in what follows, they may well be correct:

> (T2) If you judge that it's not the case that p, then you have justification to believe you don't believe that p.
> (T3) If you judge that it might not be the case that p, then you have justification to believe you don't believe that p.

justification to believe that one believes that p might well sometimes be defeated, say by opposing testimony from one's therapist. But I'll usually omit explicit reference to *prima facie* justification in order to streamline the discussion.

We should now consider how to defend the Transparency thesis. Notice that Evans says little to defend the Transparency thesis or anything like it. We can improve on his account here. In particular, we can support the Transparency thesis by using it to explain an instance of "Moore's Paradox", that much-used beast of philosophical burden.[9] To get the phenomenon in view, take a horse---"P"---and a donkey---"I don't believe that P", and conjoin them to get the following mule:

(**MP**): P and I don't believe that P.

In particular, consider judgments of contents of the form MP. Other things being equal, it is irrational make such judgments. As Goldstein (2000) would nicely put it, such judgments tend to be "Mooronic".[10]

Transparency provides a good explanation of why judgments of MP tend to be irrational. To see how, first consider that, when one judges the conjunction that [p and I don't believe that p], one judges that p, and judges that one does not believe that p. Now, if Transparency is true, one will have justification to BBp whenever one judges that p, and thus whenever one makes a judgment of the form MP. The overall upshot will be that, when one judges the MP conjunction, one judges the second conjunct, while having justification to believe the negation of the second conjunct.

As we will see in a moment, there might be cases in which your *prima facie* justification from your judgment is defeated. Nevertheless, those cases are non-standard.

---

[9] Thanks to Alan Hájek for bringing me to consider potential connections between the transparency method and "Moore's paradox".

[10] For a useful overview of forms of "Moore's Paradox", and descriptions and explanations of their absurdity, see the introduction of Green and Williams (2007).
For further discussion of Moore's paradox and issues about self-knowledge, see Shoemaker (1995) and (2009). A major divergence is that Shoemaker argues for a much stronger claim: if you believe that p (and are suitably rational and conceptually sophisticated), then you will believe that you believe that p. I suspect this claim goes too far. Consider the possibility of a thoroughgoing eliminativist with respect to belief. I take it such a character can still be rational and have the concept of belief, without ever believing she believes that p.

Your standard epistemic position is such that, if you were to judge in that position that [p and and I do not believe that p], your *prima facie* justification from your judgment would not be defeated. Therefore, if the Transparency thesis is true, it is standardly irrational to judge contents of the form MP.

Transparency also helps to capture the elusive way in which "Moore Paradoxical" judgments are distinctively defective. It is somehow defective to judge that one does not exist, but there is nothing worse about judging that [one does not exist and snow is white]. There is no tension between judging that one does not exist and judging that snow is white. To capture what is distinctively defective about Moore Paradoxical judgments, we can emphasize that judging one conjunct is in tension with judging the other, since judging the one is sufficient for having prima facie justification to reject the other.[11][12]

---

[11] My explanation has been framed using the notion of *prima facie* justification which may well be defeated. In protest, one might demand an explanation of why it is invariably irrational to make a judgment of the form MP. In response, I do not accept that there is such a general phenomenon to explain. For an interesting case in which it might not be irrational to judge a content of the form MP, consider the following remarks by Velleman and Shah:

> Arriving at the judgment that p doesn't necessarily settle the question whether one now believes it, since one may find oneself as yet unconvinced by one's own judgment. One may reason one's way to the conclusion that one's plane is not going to crash, for example, and yet find oneself still believing that it will (2003: 16-7).

We need to tweak what they say to get the crucial point. They say that one judges that not-p, yet finds oneself believing that p, but what we need is a case in which one judges that not-p while finding oneself *not* believing that not-p. Velleman and Shah's example arguably does supply such a case: one might judge that one's plane will land safely, and yet still find oneself failing to believe that one's plane will land safely. Here one certainly is not rational overall, since one appreciates the right reasons, and yet is unable to muster belief on their basis. However, one might still be justified in making the specific judgment that [my plane will land safely but I don't believe that my plane will land safely]. So it arguably sometimes is rational to make a judgment of the form MP.
For further discussion of such cases, see Gertler 2010.

[12] A further complaint would be that Transparency does not explain all versions of "Moore's Paradox", taking it that the thesis should explain all versions if it is to explain some. For further forms of "Moore's Paradox", for starters consider the (nine) combinations of the following attitudes and contents:

Attitudes: assertion, judgment, belief
Contents:
(MP1) P and I do not believe that P
(MP2) P and I might not believe that P
(MP3) I believe that P and it's not the case that P

Since the Transparency thesis explains why it is typically irrational to judge contents of the form MP, we have some reason to believe the Transparency thesis. Here we have taken an important step beyond Evans' remarks.

Let me now turn to our final main question, about *how* answering the question whether p puts you in a good epistemic position to believe that you believe that p. I will clear up two potential misunderstandings before developing my positive view.

First, it might be tempting to think that, if you form a belief that you believe that p through Evans' procedure, you do so somehow in virtue of an inference corresponding to the following argument:

p
So,
I believe that p.

The tempting thought should be resisted (but see Byrne 2005, 2010 for an important development of it). First, it jars with the phenomenology---we never find ourselves reasoning along the lines "p, so I believe that p". Second, consider the wide ranges of cases in which one can achieve knowledge through Evans' procedure (to say that there are many such cases is not to say that the procedure is invariably a source of knowledge). To put the point roughly, brains in vats or victims of Gettier cases have poorer access to the external world, but they need not have poorer access to their beliefs.

More specifically, one can know through the transparency method that one believes that p even if (a) one has a false belief that p or (b) one has a justified true belief that p while failing to know that p. Other things being equal, however, one does not achieve knowledge through inference from false lemmas or unknown lemmas. When one achieves knowledge

---

The Transparency thesis probably won't provide an explanation of what's going wrong in all of these cases (assuming that something is going wrong in all of them), but I see no reason to expect a uniform account of these potentially quite disparate phenomena.

One might still demand a uniform explanation of what's wrong with judging MP1 and of what's wrong with believing MP1. I think that even here we shouldn't expect a uniform explanation, since I take it to be worse to judge MP1 than to believe MP1.

For useful further discussion of varieties of "Moore's Paradox", and of whether they are problematic, see Hájek (2007).

through the transparency method, then, it looks like one does not do so through inference from the premise that p.[13]

We can make a similar point about the case of justification. When one has an unjustified belief that p, one can use the transparency method to achieve a justified belief that one believes that p. Other things being equal, however, one does not obtain a justified belief by reasoning from an unjustified belief.[14]

In sum, we will not understand the transparency method if we think of it as a case of somehow reasoning from the premise that p to the conclusion that one believes that p. Inference tends to obey the principle "Garbage In, Garbage Out", the transparency method does not.[15]

A second tempting thought is that, when the transparency method is a source of justification to believe that one believes that p, "inner sense" or "inner observation" is not. It is hard to make precise the notion of inner observation (although valuable clarification is available in Shoemaker 1994). Whatever inner observation may exactly be, however, one might think that the transparency method has nothing to do with it. This tempting thought is not clearly correct. A proponent of inner observation could say that, when one judges that p,

---

[13] Byrne (this volume) responds to the problem concerning knowledge by challenging the motivation of the "no false lemmas" requirement for knowledge (he does not address the corresponding problem about justification). In particular, he takes the no false lemmas requirement to be motivated by reflection on classic Gettier cases, and maintains that a better diagnosis of what has gone wrong in the cases is that the Gettierized subject fails to meet a safety requirement for knowledge. The rough idea is that the subject is in too much danger of falsely believing that p to know that p.
My challenge does not rely on a full strength necessary condition for knowledge, but instead on the heuristic that, other things being equal, one does not gain knowledge through inference from a false lemma. In any case, I doubt the safety requirement provides a better diagnosis of what has gone wrong in classic Gettier cases. Consider the original case in which Smith has a justified false belief that Jones owns a Ford, and infers the justified true belief that either Jones owns a Ford or Brown is in Barcelona. If we build into the case that it is an extremely robust fact that Brown is in Barcelona, so that Smith is safe from error with respect to the disjunction, Smith still plausibly fails to know the disjunction.

[14] According to the Evans-inspired approach of Fernández (2003, 2005) or Williams (2004), what justifies one in believing that p also justifies one in believing that one believes that p. This approach is at best incomplete, since the transparency method can give one a justified second-order belief when one lacks a justified first-order belief.
For extended critical discussion of the approach of Fernández and Williams, see Zimmermann (2004) or Vahid (2005).

[15] For further discussion of the transparency method and inference, see Gallois (1996), Brueckner (1998), and Shoemaker (this volume).

one has inner observation of one's judgment (or of the fact that one has made the judgment), and thereby has inner observation of one's belief that p (or of the fact that one believes that p). Compare: when one observes movement in a nest, one might thereby observe an object in the nest, or observe that there is an object in the nest.[16] Reflection on the transparency method might indeed remove some of the motivation for thinking that we obtain self-knowledge through inner observation. Nevertheless, it remains perfectly possible that we obtain self-knowledge through inner observation and the transparency method at the same time.[17]

I will set inner observation aside, and take up a quite different positive view about the source of one's justification when one answers the question whether p. On this view, when one gains justification for a second-order belief through the transparency method, one's judgment is itself a source of justification for the second-order belief:

> (**Transparency, Second Pass**): If you judge that p, then your judgment that p gives you justification to believe that you believe that p.

Our previous formulation of Transparency was silent about the source of one's justification for a second-order belief. The current formulation is not.

The key idea here is that judgment is itself a guide to belief. This claim is highly plausible since it explains the correlation between judging that p and having justification to believe you believe that p.

There is a significant advantage of focusing on the epistemic role of judgment in our formulation of Transparency. If one were instead to focus on the content that is judged, the transparency method could easily seem puzzling or paradoxical. Typically, when one judges that p, the content that p does not serve as evidence regarding one's beliefs. For example, that it is exactly 9:02AM is hardly evidence I believe that it is exactly 9:02AM. If one were to try to understand the transparency method by looking at the epistemic role of the content one judges, it would be puzzling how the transparency method is a source of introspective

[16] Relevant here is the "displaced perception" account discussed in Tye (2002).
[17] Contrast Richard Moran, who builds in that "a statement of one's belief about X is said to obey the Transparency Condition when the statement is made by consideration of the facts about X itself, and not by either an 'inward glance' or by observation of one's own behavior (2001:101)."

justification. If we instead focus on the role of judgment itself, however, we elegantly avoid such puzzlement about how the transparency method is a source of introspective justification. There isn't yet any puzzle or paradox of transparency.[18]

We still need to address how judgment is a source of distinctively *introspective* justification. Here it is essential to move beyond the slogan with which we began, according to which one can answer the question whether one believes that p by answering the question whether p. Given the right background information, I sometimes can answer the question whether *you* believe that p by answering the question whether p. Now I can't answer the question whether you believe you have food in your teeth by answering the question whether you have food in your teeth. However, given my background information, I can answer the question whether you believe you have gold teeth by answering the question of whether you have gold teeth. Given my background knowledge that you are unlikely to have gold teeth unawares, if I answer "yes" to the question whether you have gold teeth, I am in a good epistemic position to answer "yes" to the question whether you believe you have gold teeth. Be that as it may, I do not have introspective access to any of your beliefs.

As far as I know, there is little discussion of how the transparency method is specifically a source of introspective or "peculiar" justification. A valuable exception is Byrne (2005). To explain the peculiarity of one's access to one's beliefs through the transparency method, Byrne (2005: 96) emphasizes a difference between the following two rules:

---

[18] Contrast Byrne 2005, on whose approach a "puzzle of transparency" does arise. On his account:

this situation will be commonplace: trying to follow BEL [the rule, "if p, then believe you believe that p!"], one investigates whether p, *mistakenly* concludes that p, and thereby comes to *know* that one believes that p. (In these cases, one will know that one believes that p on the basis of no evidence at all.) (2005: 98).

It is somewhat puzzling how one could know that one believes that p on the basis of no evidence at all, but I don't think we should accept that such a situation is commonplace. Other things being equal, when one person has evidence in favor of the proposition that p, and another does not, the person with evidence is entitled to be more confident that p than the person who lacks evidence. In the case of introspection however, I take it that your confidence that you believe that p is insensitive to whether it is the case that p---it's not as if you can normally be more confident that you believe that p when you have a true belief that p as opposed to a false belief that p. So I don't think we should accept the following asymmetry required by Byrne's account: the person who uses the transparency method when it is the case that p bases their BBp on evidence, the person uses the transparency method when it is not the case that p does not base their BBp on evidence.

BEL: if p, then believe you believe that p!
BEL3: if p, then believe that Fred believes that p!

To follow a rule in Byrne's sense, one must comply with the consequent because one recognizes (and so knows, and so believes) that the antecedent is true. Therefore, if one follows BEL, one ends up with a true second-order belief---the rule is "self-verifying" in the sense that following it invariably produces a true belief. BEL3 is not self-verifying: when one follows BEL3, one may well end up with a false belief about what Fred believes.

I suspect the contrast does not explain peculiarity. Suppose (please bear with me!) that there is an omniscient God, and consider the rule:

BEL4: if p, then believe God believes that p!

If there is an omniscient God, following BEL4 will be sufficient for forming a true belief that God believes that p. Whenever one follows BEL4 in Byrne's sense of "follow", one recognizes that p, and so knows that p, so it's the case that p. Therefore, whenever one follows BEL4 in Byrne's sense of "follow", the omniscient God will know that p, and so the omniscient God will believe that p. However, despite all this, one still does not have properly introspective access to the beliefs of the omniscient God.

To ensure that judgment is a source of distinctively introspective justification, I will say that one does not rely on background information when one gains justification from one's judgment to believe one believes that p. This secures a contrast between the first person case and the third person case, as well as being phenomenologically plausible. I will thus refine our characterization as follows:

> (**Transparency, Third Pass**): If you judge that p, then your judgment that p gives you immediate justification to believe that you believe that p.[19]

[19] Here I do not assume that all introspective justification is immediate (we will see cases of inferential introspective justification in what follows).
I should say that there are alternative explanations of the contrast. For example, one might say that you rely in your own case on the background belief that, if you judge that p, then you believe that p, whereas in the case of others you rely on the background belief that if p, then the other person believes that p. I think this explanation is worse than my own, on the grounds that it over-intellectualizes the transparency method. In particular, the transparency method is available in a wider range of cases than those handled by the alternative

I will make one more point about the way in which judgment is a guide to belief. Judgment is a fallible guide to belief. To adapt what Descartes said in the Discourse on Method, the action of thought by which we gain justification for a second-order belief can occur in the absence of the relevant first-order belief. For an important type of case, consider the following: you judge that your flight leaves at noon, and then realize that you do not and did not believe that your flight leaves at noon. In such cases, your judgment that p is a kind of performance error which fails to reflect an underlying belief---"what was I thinking?", you might go on to say. You "blurted out" that p, either in speech or in merely in thought, consciously endorsing the proposition that p, yet failing to have a standing belief that p.

Judging that p is insufficient for believing that p, I take it, because believing that p requires having various dispositions, where judging that p is insufficient for having those dispositions.

In protest, one might say that judgment that p is a species of conscious belief that p, so that it is impossible to judge that p without believing that p. This objection misses the point. Our focus is on the way in which judgment is a guide to what we might think of as standing beliefs, or as dispositional beliefs.[20] For the purposes of the paper I have reserved the term "belief" for such states. As far as the focus of this paper is concerned, it doesn't matter whether there is a wider use of the term "belief" which encompasses judgments themselves.[21]

---

explanation. The first type of cases involves those who have the concept of belief but not yet the concept of judgment: such people may use the transparency method but do not have the background belief that if you judge that p then you believe that p. The second type of case involves those who have the concept of belief and the concept of judgment, but who are (for whatever bizarre reason) thorough eliminativists about judgment although not eliminativists about belief. Such people may also use the transparency method to acquire a justified belief they believe that p, but presumably won't be relying on a background belief that if they judge that p then they believe that p.

[20] A separate question concerns how judgment might itself be a guide to judgment. It might turn out that judgments are an infallible guide to judgments, but my focus is on how judgments are a fallible guide to beliefs.

[21] Still, there is controversy about whether one can judge that p without having a standing belief that p. Although the affirmative view is defended for example by Peacocke (1999), it is denied for example by Zimmerman (2006). I should emphasize that, even if judging that p did suffice for believing that p, it would be enough for my purposes if there is some state phenomenologically just like judging that p which can occur in the absence of belief. That's because, in the cases in which one is in a judgment-like state without believing that p, I will hold that the judgment-like state still gives one immediate fallible justification to believe that

Also, one might protest that, if a judgment gives one fallible justification for a second-order belief, then it cannot give one immediate justification for a second-order belief. On this line of thought, if a state gives one fallible justification to believe that p, then there is a "gap" between the obtaining of the state and the truth of the proposition that p, so that some intermediate belief will be needed so as to "bridge the gap" between them. However, if the intermediate belief is itself only fallibly justified, the line of thought can be repeated concerning its own justification. And since there are few infallibly justified beliefs, the line of thought threatens to show that there are few fallibly justified beliefs. Since many of our beliefs are fallibly justified if justified at all, the reasoning threatens to have skeptical consequences. We should not endorse it.

We may now formulate the Transparency thesis as follows:

> (**Transparency**): If you judge that p, then your judgment that p gives you immediate fallible justification to believe that you believe that p.

This is the lesson to take away from reflection on Evans' classic passage.[22]

2. The Case Against Transparency

---

one believes that p. Although the fallback position is available, it is simpler to work with talk about judgment.

Zimmerman (2006) objects to the fallback position as follows:

> . . . if experience with the phenomenal character of genuine judgment is not sufficient for belief, we can have experiences with this phenomenal character that are not real judgments (for they don't initiate, sustain or accompany beliefs). If our second-order introspective beliefs are grounded in such judgment-like experiences, knowledge of our beliefs is not direct, but instead mediated by inconclusive inferential grounds or states of inner perception (367).

In response, I see no reason to accept the dilemma he proposes. Judgment-like experiences arguably can provide fallible immediate justification, we need an argument that they can't. (I will say much more about fallible immediate justification in what follows).

[22] For a different take on the transparency method, in terms of considerations about "rule-following", see Byrne (2005) and Setiya (forthcoming). For criticisms of Byrne (2005) see Shoemaker (this volume).

A question I leave open is whether the transparency method can somehow be generalized beyond the case of belief. For discussion of the case of visual experience, see Evans (1982), Peacocke (2008), and Byrne (2005). For discussion of further cases, see Gordon (1995, 2007), Byrne (2005), and Way (2007).

I will now consider key objections to the Transparency thesis. I will raise two challenges briefly before moving on to my main discussion.

2.1. Preliminary Challenges

The first objection uses a constraint on immediate justification:

> (**Face Value Constraint**): Necessarily, if a state M gives one immediate justification to believe that p, then M has the content that p.

I grant that the Face Value Constraint is attractive. In particular, it meshes well with the plausible claim that, when one forms an immediately justified belief on the basis of a state, one does so by taking some content of the state at face value. Given that a judgment *that p* does not have the content that *one believes that p*, the Face Value Constraint predicts that judgments that p are not sources of immediate justification for self-ascriptions of beliefs that p.

As attractive as it is, the Face Value Constraint is false. One way to see this is by considering the case of consciousness. Your state of being conscious can give you immediate justification to believe that you are conscious, whether or not you are in any state with the content that you are conscious.[23] A somewhat more controversial case to consider is that of belief. Arguably one's belief that p can itself give one immediate justification for a second-order belief (Zimmermann 2006). If that's right, then the Face Value Constraint is false, since one's belief that snow is white does not have the content that one believes that snow is white, the first-order belief instead simply has the content that snow is white.

The challenge from the Face Value constraint fails.[24]

---

[23] According to philosophers such as Searle (1983), visual experiences have contents which involve references to the experiences themselves. This is not yet to say that any experience of mine will have the content that *I* am conscious.

[24] For a nice discussion of related issues, see Pryor (2005).
One might wonder whether there are counterexamples to the Face Value Constraint which are not cases of introspective justification. I believe such cases arise in the epistemology of perception, and discuss them further in (forthcominga).

The next objection is inspired by a difficulty which arises in the epistemology of perception. The worry in the perception case is that, if our experiences do provide immediate justification for external world beliefs, then it will be too easy for us to reject skeptical hypotheses about our experiences. In particular, we might be able to justifiedly reject skeptical hypotheses simply by performing inferences which correspond to the following argument:

I have hands.
If I have hands, then I am not a handless brain in a vat.
So,
I am not a handless brain in vat.

According to the challenge, since we do not gain justification to reject skeptical hypotheses by performing such inferences, we do not gain immediate justification from our experiences for external world beliefs either (Cohen 2002, 2005, Wright 2002

There is a parallel objection to my account of how judgment is a guide to belief. The accusation would be that, if judgment is a fallible source of immediate justification, then it will be too easy for one to gain justification to reject skeptical hypotheses which say I judged that p without believing that p. Since it is in fact not so easy for one to gain justification to reject those skeptical hypotheses, judgment is not a source of immediate justification after all.

In both cases, the idea is that we lack basic justification since we lack easy justification to reject skeptical hypotheses.

In brief, I think the best reply separates issues about anti-skeptical justification from issues about introspective justification. We can enjoy basic justification for our introspective beliefs whether or not we enjoy easy justification to reject skeptical hypotheses. But it takes a paper or two to properly develop the line of objection and explain why it fails. I do the needed work in my (2008) and (forthcomingb).

2.2. The Constitutivist Critique

I will now turn to the main challenge I will address in this paper.

According to the moderate position I have developed, what justifies the subject's second-order belief in some cases is compatible with the falsehood of the second-order belief. However, one might think I have underestimated infallible sources of justification.

The line of objection is based on so-called "constitutivist" views about introspection, so-called because they propose a connection between the nature of belief and second-order beliefs. The approach has been taken up in very different ways by philosophers such as Sydney Shoemaker (1996, 2009), Richard Moran (2001), Jane Heal (2002), and Akeel Bilgrami (2006).

Since the line of objection raises more general issues about what it is to believe that p, it is of much wider interest than just as a challenge to my view.

The approach's most prominent proponent, Sydney Shoemaker, develops the approach by focusing on the relation between first order beliefs and second order beliefs. In order to articulate a view specifically about the *epistemology* of introspection, however, we need to look at the relation between first order beliefs and epistemic states such as justified second order beliefs or knowledge. Given that there are many epistemic states, there are also many ways to develop the constitutivist position.

My aim is to build the view from the ground up as a view in epistemology. I will therefore depart from what promising possibilities there are in logical space, rather than from quotations from key figures (although their remarks will still play a role in guiding our development of the view). I will also bring the view to engage with discussions in epistemology in general, rather than only in the philosophy of mind in particular.

We will survey a number of constitutivist claims in what follows. I will argue that the stronger claims invoked are false or unmotivated, and the weaker claims invoked are compatible with my own position.

The constitutivist must address several choice points. The first I will discuss is whether the view will concern an entailment from beliefs to epistemic states, or instead from epistemic states to beliefs. Let's start by considering entailments from beliefs to epistemic states, and in particular the epistemic state of knowledge. Here we are concerned with the manner in which beliefs might be "self-intimating".

2.2.1. Self-Intimation

A very demanding starting point would be the claim that

> (**Knowledge**): Necessarily, if S believes that p, then S has introspective knowledge that she believes that p.

This proposal is far too strong. Assuming that knowledge entails belief, the proposal straightaway requires an infinite regress of higher order beliefs. Since the regress does not look benign, we should use a less demanding proposal.[25]

A weaker view in terms of knowledge is that

> (**Potential Knowledge**): Necessarily, if S believes that p, then S is in a position to have introspective knowledge that she believes that p.

Roughly speaking, one is in a position to know that p when one's epistemic position is good enough for one to know that p, although one may not yet have met the psychological requirements for knowing that p, such as that of believing that p. Given that the current proposal abstracts from psychological requirements, it avoids the regress generated by the first.[26]

The proposal is still extremely demanding. I think the constitutivist should avoid commitment to it.

A specific reason to do so is supplied by Williamson (2000)'s powerful anti-luminosity argument. It's worth reviewing the argument to appreciate the challenge. On closer scrutiny, Williamson's argument actually indicates that my fallibilist view about introspective justification is correct. So in the course of considering how to formulate a constitutivist view about self-intimation, we'll actually see further evidence in favor of my fallibilist approach to introspection.

---

[25] Although see Shoemaker (2009: 42-43) for an argument that a similar regress is in fact benign.

[26] Another way to proceed is by packing psychological requirements into the antecedent of the formulation of the view. These sorts of formulations would proceed as follows:
If one believes that p, AND considers the question whether p . . .
As far as I can tell, there will be no substantive difference between this family of formulations and the one I consider in the main text. One might wonder which family of formulations is more fundamental, assuming that one is, but I won't take on this question.

The setup of the argument is as follows. Consider a series of times during which Belle's confidence that p very gradually decreases, so that she believes that p at the beginning but does not believe that p at the end. Throughout the series Belle carefully considers whether she believes that p, so that whenever she is in a position to know she believes that p, she does know she believes that p. Finally, her confidence that she believes that p likewise gradually decreases throughout the series.

The crucial premise of the argument is that, if Belle knows at a given moment in the series that she believes that p, then she believes that p at the next moment in the series. To see why the premise is plausible, suppose it is false. If it is false, then there is some moment in the series at which Belle knows that she believes that p, and an adjacent moment at which she does not believe that p. This would require that she knows a proposition at the earlier moment despite having misplaced confidence at the next, where that misplaced confidence is at a very close level on a very similar basis. But knowledge plausibly requires the avoidance of such forms of error---there is plausibly a "safety" requirement on knowledge. So the crucial premise is plausibly correct.[27]

Let's now consider how to argue from the crucial premise to the falsehood of Potential Knowledge. At the first moment in the series, Belle knows that she believes that p. By the crucial premise, it follows that, at the second moment in the series, she does believe that p. By Potential Knowledge, at the second moment in the series, she is in a position to know (introspectively) that she believes that p. By the setup concerning her attentiveness, at the second moment in the series she does actually know that she believes that p. This reasoning can be repeated until we reach the conclusion that Belle believes that p at the last time in the series. Something has gone wrong!

---

[27] For some discussion of how to clarify the safety idea, see Sosa (1999), Pryor (2004), and Manley (2007). For some criticism, see Neta and Rohrbaugh (2004) or Comesaña (2005). Notice that given the setup of the anti-luminosity argument, it is unable to challenge one family of extremely strong claims about self-knowledge. Consider "revelation" theses about mental states, according to which one is in a position to know the *essence* of a mental state by being in it. Since one's being in a given mental state does not belong to its essence, and more generally since the essence of a mental state doesn't change over time, Williamson's argument won't threaten any revelation thesis.
For further discussion of revelation theses about mental states, see Lewis (1995) and Stoljar (2009).

One might object to the crucial premise, or even to some aspect of the setup of Williamson's argument, but I think the constitutivist is best advised to avoid engagement with it altogether.[28] That's because there's a much more general reason to doubt the knowledge proposals: our epistemic humility should extend to propositions about our own minds and not just to about the world. Just as we shouldn't infer that a worldly proposition is false from the fact that we are in no position to know it, we shouldn't infer that a belief-ascribing proposition is false from the fact that we are in no position to know it.

This humble line of thought does not beg the question against constitutivism. Constitutivists might want to say that our access to our own minds is superior to our access to the world, but we can comfortably accept this thought without going to the extremes of the knowledge proposal.

The constitutivist view should be formulated in terms of some less demanding epistemic position than knowledge. One way to try to do this is as follows:

> (**Reason**): If you believe that p, then one of the reasons you have to believe you believe that p is that you believe that p.[29]

It is tempting to think that this view is a cautious alternative to the earlier claims. This tempting thought is not clearly correct (it is also not clearly incorrect). To see how the complication arises, we should ask what it takes for one to have the reason that p to believe that q. According to the knowledge-theoretic tradition defended by Unger (1975), Williamson (2000) and Hyman (1999),

> (**Reason-Knowledge**): You have the reason that p to believe that q only if you know that p.

If this tradition is correct, the Reason thesis entails the extremely strong Knowledge thesis, and is hardly any alternative to it.

---

[28] Useful critical discussions of the argument include those by Conee (2005) and Berker (2008).

[29] Shoemaker and Zimmermann might be inclined to go along those lines. For instance, Shoemaker writes that "belief in a proposition *provides a reason to believe* a proposition---the normative proposition that one ought to be guided by that proposition in one's thought and action---which is arguably coextensive with the proposition that one believes that proposition (2009: 39, emphasis mine).

We should avoid taking a stand here on what reasons are, and what it takes to have them. To avoid the complication, the constitutivist shouldn't present her view specifically in terms of reasons, but instead more generally in terms of sources of justification. Consider that an experience can give a child justification for a belief, even if the child lacks the concepts required to know that she has the experience. Even if a knowledge requirement holds for reasons, there is still room for a more relaxed view about other sources of justification than reasons.[30] The constitutivist is therefore better advised to work with the following proposal:

> (**Justification**): Necessarily, if you believe that p, then your belief that p gives you introspective justification to believe you believe that p.[31]

This proposal more clearly avoids Williamson's anti-luminosity argument. Here is a first approximation of the reason why. Although it is plausible that knowledge requires the avoidance of error in nearby cases, it is not plausible that justification requires the avoidance of error in nearby cases. If there were such a requirement on justification, it would not be possible to have justification for a false belief, given that a false belief trivially involves error in a nearby case. I assume however that it is possible to have justification for a false belief.[32]

For this response to the anti-luminosity argument to be effective, it actually needs to be put in terms of introspective justification in particular rather than in terms of justification in general. Otherwise the threat would remain that there is a safety requirement for introspective justification, even though there is not a safety requirement for justification. However, if one adapts the reasoning in the previous paragraph all the way through, one will directly commit oneself to my view that one can have introspective justification for a false second-order belief. But then the constitutivist will be committed to one of my key claims.

---

[30] One might of course not be so relaxed. For discussion of how to push the knowledge-theoretic approach further, see Williamson (2000).

[31] I should say that I have in mind justification for outright belief as opposed to for a mere increase in confidence. There are less demanding formulations of constitutivism but I won't pursue them further here.
For further discussion of various fallback position one might adopt in response to the anti-luminosity argument, see Greenough (this volume) and Smithies (this volume).

[32] This point can also be found in Conee (2005) and Berker (2008).

The more general strategy of the response to Williamson is to distinguish between what it takes to have introspective knowledge and what it takes to have introspective justification. So long as the constitutivist pursues this general strategy, it's not clear how she will be entitled to reject my view that one can have introspective justification for a false second-order belief. Given that introspective justification is unlike knowledge in that it lacks a safety requirement, it is not clear why introspective justification should be similar to knowledge in being factive. If introspective justification is compatible with the nearby possibility of error, it's unclear why, in the setup of Williamson's argument, one shouldn't retain introspective justification beyond the last moment at which one believes that p.[33]

In sum, to avoid the anti-luminosity argument, one should contrast what it takes to have introspective justification from what it takes to have knowledge. To make this move is to play into my hands. The less introspective justification looks like knowledge, the more it looks like it should be possible to have introspective justification for a false belief.

Let me now return to the more general question of how to understand the constitutivist position.

The formulations just given are all in terms of entailment. A further choice point for the constitutivist is whether to put their proposal in stronger terms, in particular in terms of the essence of belief rather than merely in terms of an entailment from belief. The two ideas are different. There is an entailment from my belief that snow is white to the fact that 2+2=4. However, the essence or nature of my belief is more fine-grained, and has nothing to do with the fact that 2+2=4. Mathematical facts are neither here nor there when it comes to acquiring a full understanding of what it is to believe that snow is white.[34]

The constitutivist is well advised to explain the necessity involved in Justification, rather than to leave it unexplained. A good way to do so, although not the only way to do so, is to endorse the following stronger claim:

---

[33] One might respond by making the further claim that a belief that p is the *only* potential source of introspective justification to BBp. Such a position of course needs further defense, but it would explain why one might have introspective justification to BBp only if one believes that p. Making this move would commit the constitutivist to the following position: One has introspective justification to believe that p if and only if one believes that p. I give critical discussion of the position below.

[34] For valuable discussion of essence vs. necessity, see Fine (1994).

(**Essential Justification**): If you believe that p, then it is of the essence of your belief that p to give you introspective justification to believe you believe that p.

By focusing on the nature of beliefs, rather than merely on what they entail, I think we have moved closer to the heart of the constitutivist approach.[35]

I set aside the further evaluation of Justification or Essential Justification. I do so because both of these claims are actually compatible with our account of the transparency method. Their main upshot is to give beliefs a role in the epistemology of introspection. They do not say anything, or clearly imply anything, about the role of conscious judgments in the epistemology of introspection. To include beliefs is by no means to exclude judgments. Moreover, the proposals do not say or clearly imply anything about whether there is fallible introspective justification. The formulations are restricted to cases in which one does believe that p. They are silent about cases in which one does not believe that p.

I hope so far to have improved our understanding of the options open to the constitutivist. As the view has been developed so far, it has turned out to be no threat to my own. To isolate the threat to our treatment of transparency, we need to look at an entirely different family of constitutivist views.

2.2.2. Infallibility

To isolate a challenge to my view, we need to look at claims which at least concern an entailment from epistemic states to first-order beliefs, rather than entailments from first-order beliefs to epistemic states.

---

[35] Sometimes the constitutivist view is put in terms of the essence of rationality rather than in terms of the essence of belief. Consider the following comment by Shoemaker 2009:
"this seems a step towards the view that beliefs are constitutively self-intimating---that it is part of being a rational subject that belief that p, together with the possession of the concept of belief and the concept of oneself, brings with it the belief that one believes that p (2009: 36)."
If *beliefs* are to be constitutively self-intimating, I think we should focus on what it is to believe that p rather than on what it is to be rational. In particular I think we should instead say that
"it is part of believing that p that, if one believes that p, and is rational and has the concepts of belief and oneself, then one believes that one believes that p".

As before, the epistemic state in question might be that of knowledge or that of justification (setting aside even stronger or even weaker states).

I will focus on the case of justification, and will start with the following proposal:[36]

> (**Justification2**): If one has introspective justification to believe one believes that p, then one believes that p.[37]

This thesis concerns introspective justification, but it does not say anything further about sources of introspective justification. Given this omission, Justification is compatible with the following more specific proposal:

> (**Observational Justification**): Necessarily, if one has introspective justification to believe one believes that p, then one has inner observation of one's belief that p.

The current proposal takes the stand that inner observation is a source of introspective justification. But this idea is standardly odious to the constitutivist. We need to look further to capture the spirit of their view. We need at least the following more specific claim:

> (**Constitutive Justification**): If you have introspective justification to believe you believe that p, then what gives you justification for the second-order belief is that you believe that p (rather than inner observation of your belief that p).

---

[36] The knowledge case is less controversial. Since almost everyone agrees that one can know a proposition only if it is true, almost everyone will agree at least that

Necessarily, if one knows that one believes that p, then one believes that p.

There are further questions to address about the ground of one's knowledge, which are so far left open, but I will set them aside. So long as one knows only true propositions, thinking about knowledge will not tell us whether one can be in a good epistemic position through introspection for a false belief.

[37] This claim is endorsed in Zimmerman (2006, section 8). Consider his remark that "when our second-order introspective beliefs are formed and maintained in a first-person way they are grounded in the very first-order mental states that make them true (370)." However, he sometimes presents his view in a more qualified way: "if we have any false, justified beliefs about what we believe, the grounds for these beliefs will be different in kind from the grounds with which we hold our typical second-order introspective beliefs (371)." One could accept this quote while allowing for an introspectively justified yet false second-order belief.

The claim has very striking consequences. If it is true, then one cannot have a second-order belief which is both false and introspectively justified. As far as second-order beliefs are concerned, then, CJ embodies an infallibilist conception of introspective justification. Also, the thesis requires that *only* beliefs provide introspective justification for second-order beliefs. So CJ rules out judgment is a source of justification at all.

The Constitutive Justification claim is striking but mistaken. We can have fallible introspective justification for our second-order beliefs, just as we can have fallible justification for other beliefs. The most convincing way to see why is by considering cases of inferential introspective justification. I will set out such a case in some detail, since it provides a principled way to argue against CJ.

Consider a quite idealized subject: when something she believes has a strictly logical consequence, she reliably tends to believe the consequence. She also has justification to believe that she is thorough in this way. However, she is not perfect, since she does not always follow through with the logical consequences of what she believes, and she also sometimes makes justified mistakes about what is a strictly logical consequence of what. Now suppose she reasons in a quite indirect way about what she believes, in a way she would vocalize as follows:

"(1) I believe that p.
(2) As a matter of logic, if p, then q.
(3) If [I believe that p and, as a matter of logic, if p, then q], then I believe that q.
So,
(4) I believe that q."

Focus on cases where she has a justified false belief in (2)---I take it we can all make such mistakes when logic gets hard. In such cases she can still have an introspectively justified belief in (1), and a justified belief in (3). As long as these pieces are in place, she presumably will end up with a justified belief as well in (4), given that it is an obvious consequence of the contents of other justified beliefs she has. Further, her belief in (4) should be introspectively justified in particular.

All of these pieces can be in place whether or not (4) is true. When she reasonably misidentifies a logical entailment, she can fail to believe the proposition she takes to be a strict logical consequence of another proposition she believes. Remember that in doing the

reasoning she would vocalize with (1) through (4), she is forming a belief she believes that q, and need not be forming a first-order belief that q. In particular, as far as her explicit reasoning is concerned, she is not reasoning from her belief that p to form a belief that q. Now, her lack of a belief that q need not interfere with her justification to believe any of (1) through (3). So her belief in the conclusion will be justified---introspectively---whether or not it is true. Given that this sort of case is possible, the Constitutive Justification thesis is false.

In response, someone might say that the subject is not justified at all in believing (4). Whether or not there are counterexamples to the principle that justified belief is closed under obvious consequence, I take it that the current example is not such a case.[38]

In a separate response, someone might say that the subject is not *introspectively* justified in believing the conclusion, on the grounds that she is inferentially justified in believing the conclusion.[39] But something can both be an inferential and an introspective source of justification. For example, it might be through reasoning about counterfactual situations that you realize you hope that p, rather than expect that p (Williamson 2000). In such a case you have inferential justification to believe you hope that p, but the case is still a paradigm of introspective justification, since no one else can gain justification to believe you hope that p in the way you did.

---

[38] Shoemaker (this volume) argues that it is not even possible to falsely believe one believes that p. In particular, he writes that

> Will it [the second order belief] bestow the disposition to assent to the content of that putative belief? If it does, we will then have a case for saying that the person does believe that content, or at least that it is not determinately true that he does not. If it does not bestow that disposition, then the person will be liable to fall into a version of Moore's paradox---saying, or thinking, "I believe that p, but not-p," or "I believe that p, but I have no idea whether p is true (44).

In response, if a subject with a false second-order belief must fall into a version of Moore's paradox, I would simply accept that this can happen. Being subject to Moore's paradox is no barrier to having a false second-order belief, or even to having an introspectively justified false second-order belief. We can agree that it is irrational to believe the relevant conjunction, while still maintaining that it is possible to believe the relevant conjunction. In particular, it could even be that the best explanation of why it is irrational to believe the conjunction proceeds in terms of one's having introspective justification to believe the first conjunct.

[39] See e.g. Fernández (2005: 541-2):

> Whatever adopting that [first-person] perspective on our own beliefs ultimately amounts to, it is a way of forming beliefs about them that provides one's meta-beliefs with a special kind of justification, in that:
>
> (i) It does not depend on reasoning.
>
> (ii) It does not depend on behavioral evidence.

One might have thought that all introspective justification is immediate justification. This thought is wrong.

The case I just presented only concerns inferentially justified beliefs. The proponent of CJ therefore might fall back to the following weaker claim:

> (**Constitutive Justification 2**): Necessarily, if you have immediate introspective justification to believe that you believe that p, then what gives you justification for the second-order belief is that you believe that p.

Since CJ2 is merely concerned with immediate introspective justification, it is silent about the type of case just discussed. The claim is still striking. If it is true, then one cannot have a false second-order belief which enjoys immediate introspective justification. So CJ2 embodies an infallibilist conception of immediate justification, at least as far as second-order beliefs are concerned. Next, if CJ2 is true, *only* beliefs provide immediate introspective justification for second-order beliefs. The thesis thereby rules out that judgments also play this role. So the thesis is still a threat to my overall position.[40][41]

I am not aware of any uncontroversial counterexamples to CJ2. I take the transparency method to provide counterexamples to CJ2, but these cases will not be convincing to the proponent of the claim. The point I will like to press is that CJ2 needs to be defended, where it's far from clear how to defend the claim.

There is an interesting and important way to emphasize that CJ2 needs special defense. Consider its converse:

---

[40] There is a different way to introduce the alternative formulation (for a related point see Smithies, this volume, section 4). One might say that Constitutivism concerns only "pure" or "wholly" introspective justification, where all pure introspective justification is immediate. Perhaps my earlier cases failed to target the view as it is properly understood, since the cases involved a mixture of introspective and non-introspective justification. However, this response involves the constitutivist in unnecessary controversy. When I realize that I hope that p through counterfactual reasoning, I arguably have pure introspective justification, while still failing to have immediate introspective justification. Again, we should not take the case of pain as the only paradigm.

[41] I should mention that, in assigning judgments a role in our access to our beliefs, I do not yet assert that judgments play the central role in our access to our beliefs. For useful discussion of this issue, see Zimmerman (2006), Shoemaker (2009: section 10), and Smithies (this volume).

(**Converse**): Necessarily, if you believe that p, then you have immediate introspective justification to believe you believe that p.

The constitutivist could explain why this formulation is true by making a further claim about the essence of belief. CJ2 is importantly different however. To give a parallel explanation, the constitutivist would need to invoke a claim about the essence of non-belief, to the effect that the essence of non-belief involves access to the absence of belief. But such claims are mistaken. My chair doesn't have any belief, and also doesn't have any reason to think that it lacks beliefs. It does not have any mental states or justifications at all. Whether or not belief has a (partly) epistemic essence, non-belief does not have an (even partly) epistemic essence. In any case, there is no need to give a special account of the nature of the absence of belief. In general, the essence of the absence of x is just the absence of the essence of x.[42]

Given the asymmetry between the two proposals, I take CJ2 to be in special need of defense. I take the spirit of constitutivism to concern an epistemic claim about the nature of belief. Given that no such claim holds about non-belief, there is a special concern about how to motivate CJ2. Given that there is no account of the essence of non-belief in terms of access to non-belief, the thesis is arguably not in the spirit of constitutivism, and hence would presumably need to be defended on very different grounds.

One way to defend CJ2 would be to say that, in general, immediate justification is infallible. Call this the infallibilist conception of immediate justification. Even if this demanding view is true, it doesn't quite get us to CJ2. The infallibilist conception tells us that every immediately justified second-order belief is true, but it doesn't tell us what the source is of their justification. Claims about the source of a justification are quite different from claims about the infallibility of a justification. For example, an immediately justified belief that 0=0 is trivially an infallibly justified belief, but it's dubious that the fact that 0=0 is somehow itself the source of the justification of the belief. Since CJ2 does make a further claim about the source of the immediate justification for second-order beliefs, CJ2 doesn't obviously follow from the infallibilist conception of immediate justification.

---

[42] One might try to explain CJ2 by instead making the following claim about the essence of belief: it is of the essence of believing that p that *only* a belief that p can give one immediate introspective justification to believe that one believes that p. I don't have any sophisticated objection to this proposal, my main concern is that it seems contrived.
For further discussion of cases of absences of belief, see Sosa (2003)

The challenge for the constitutivist here is serious. She needs to distinguish her position from the odious inner observation theories. She needs to take a stand on the source of introspective justification. Since the odious inner observation theory could take on board infallibilism about immediate justification, the constitutivist needs a more discriminating defense of her position.

A major further problem is that there is no reason to believe the infallibilist conception. As I emphasized in the introduction, immediate justification is not characterized by its strength, but instead by its lack of dependence on one's justification for background beliefs. Given that immediate justification is not characterized by its strength, there's no reason to expect it to be infallible. CJ2 needs special pleading in its defense.

The most promising argument for CJ2 I am aware of returns to a "self-intimation" thesis.[43] The argument relies in particular on the idea that *absences* of belief are strongly self-intimating, as well as the idea that you cannot have justification to believe each of two contradictory propositions at the same time:

(H) If one does not believe that p, then one has immediate introspective justification to believe one does not believe that p.
(I) If one has justification of any kind to believe that p, then one does not have justification of any kind to believe it's not the case that p.
So,
(J) If one does not believe that p, then one does not have immediate introspective justification to believe one believes that p.

The conclusion of this argument is equivalent to the claim that, if one does have immediate introspective justification to BBp, then one does believe that p. The conclusion is thus quite close to CJ2, although CJ2 does make the further claim that one's belief that p is the *source* of one's introspective justification.

Setting aside the problem that an inner observation view has yet to be excluded, the main problem with the argument is that it equivocates. For the claims about self-intimation to be correct, they must be read in terms of *prima facie* justification. When one seems to not have a belief, it is still possible to gain evidence that one has the belief, say from one's therapist. For the crucial claim about incompatibility to be correct, however, it must be understood in terms of *all things considered* justification---there is no difficulty in having

[43] Conversations with Declan Smithies and Daniel Stoljar were extremely helpful here.

prima facie justification for each of two contradictory claims. The argument is therefore invalid when the premises are construed in the way in which they might be true.

I am aware of no better way to defend the view that introspective justification is infallible. The constitutivist challenge to our moderate view does not succeed.

Conclusion

Although there is much disagreement about self-knowledge, philosophers in the literature currently tend to agree that it should not be understood on the model of perception. They focus on the metaphysics of perceptual states, and on the absence of appropriately similar states in the case of introspection. The idea is that there is no good sense in which we perceive our mental states.

The current focus on metaphysics has obscured parallels between the *epistemology* of perception and the epistemology of introspection---in each case a state can give a kind of justification which is immediate yet fallible. Just as an experience can give one immediate justification to believe that p, even though one can have the experience when it's not the case that p, a judgment can give one immediate justification to believe that one believes that p, even though one can make the judgment without believing that p. That is the proper understanding of the poorly understood "transparency of belief". Judgments play such a role in our introspective lives, not all the work can be done by beliefs themselves.

I would hold that there is a non-inferential yet fallible structure in many other cases of introspection. Consider our access to factive mental states such that of seeing that p or remembering that p, where one can be in such mental states only if it is the case that p. Or consider our access to relational mental states such as that of seeing o, where one sees o only if one is appropriately interacting with the thing. In each of these cases, we need to be able to account for beliefs which are first-personally justified, yet false. To get a good view of the epistemology of introspection, then, we should not look away from the epistemology of perception.

In developing my account of the transparency method, I have emphasized the role of conscious judgment in our introspective lives. Doing this work should improve our understanding of consciousness itself. Let me briefly sketch two potential upshots. First consider the epiphenomenalist view that consciousness plays no causal role. In order to form

a second-order order belief on the basis of a conscious judgment, however, it looks like the belief must be caused or causally sustained by the judgment. We thus have a new reason to avoid epiphenomenalist views of consciousness---consciousness must play a causal role to play its epistemic role. Second, our work should also inform current debate about the phenomenology of cognition.[44] What is it like, if anything, to think that p? Given that different conscious judgments can justify us in self-attributing different beliefs, we might expect judgments with different contents to have different conscious characters. If what it's like to judge that p were the same as what it's like to judge the different proposition that q, it would be unclear how the judgments could still differ with respect to which self-attributions they justify. To see the point, consider the following (imperfect) analogy: if what it's like to see redness were the same as what it's like to see greenness, it would be unclear how the color experiences could differ with respect to which color attributions they justify. Our work therefore suggests there is some support to views on which the phenomenology of cognition is fairly rich. To develop and assess the argument in detail is a further matter.

References

Berker, S. 2008: "Luminosity Regained" *Philosophers' Imprint*, 8: 1-22.
Bilgrami, A. 2006: *Self-Knowledge and Resentment*. Cambridge: Harvard University Press.
Boyle, M. 2009: "Two Kinds of Self-Knowledge", *Philosophy and Phenomenological Research*, 78: 133-64.
Brueckner, A. 1988: "Moore Inferences", *Philosophical Quarterly*, 48: 366-9.
Burge, T. 1996: Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society* 96:91-116
Byrne, A. 2005: "Introspection", Philosophical Topics, 33: 79-104.
Byrne, A. 2010: "Knowing that I am Thinking" in Anthony Hatzimoysis (ed.) *Self Knowledge*. OUP.
Cassam, C. 2007: *The Possibility of Knowledge*. Oxford: OUP.
Chalmers, D. 2003: "The Content and Epistemology of Phenomenal Belief", in Q. Smith and A. Jokic (eds), *Consciousness: New Philosophical Perspectives*. Oxford: OUP.
Cohen, S. 2002: "Basic Knowledge and the Problem of Easy Knowledge", *Philosophy and Phenomenological Research*, 65: 309-29.
----- 2005: "Why Basic Knowledge is Easy Knowledge", *Philosophy and Phenomenological Research*, 70: 417-430.
Conee, E. 2005: "The Comforts of Home" *Philosophy and Phenomenological Research*, 70: 444-451.
Davies, M. 2004: "Epistemic Entitlement, Warrant Transmission and Easy Knowledge",

[44] For some discussion, see Siewart (1998) or Pitt (2004).

Aristotelian Society Supplementary Volume, 78: 213-45.
Edgley, 1969: *Reason in Theory and Practice*. London: Hutchinson.
Evans, G. 1982. *The Varieties of Reference*. Oxford: OUP.
Fernández, J. 2003: "Privileged Access Naturalized", *Philosophical Quarterly*, 53: 352-372.
----- 2005: "Self-Knowledge, Rationality, and Moore's Paradox", *Philosophy and Phenomenological Research*, 71: 533-556.
Fine, K. 1994: "Essence and Modality", *Philosophical Perspectives*
Gallois, A. 1996. *The World Without the Mind Within*. Cambridge: CUP.
Gertler, B. 2010, "Self-Knowledge and the Transparency of Belief" in Anthony Hatzimoysis (ed.) *Self Knowledge*. OUP.
Goldstein, L. (2000), 'Moore's Paradox', in P. Engel (ed.), *Believing and Accepting* (Dordrecht: Kluwer Academic Publishers), 65 – 92.
Gordon, R. 1995: 'Simulation without Introspection or Inference from Me to You', in *Mental Simulation: Evaluations and Applications*, eds. T. Stone and M. Davies (Oxford: Blackwell, 1995), 53–67.
----- 2007: "Ascent Routines for Propositional Attitudes", *Synthese* 159:151-65.
Green, M. and Williams, J. 2007. *Moore's Paradox: New Essays on Belief, Rationality, and the First Person*. Oxford: OUP.
Hájek, A. (2007): "My Philosophical Position Says 'p' and I Don't Believe 'p'. In Green, M. and Williams, J. (eds) *Moore's Paradox: New Essays on Belief, Rationality, and the First Person*. Oxford: OUP.
Hazlett, A. 2010: "The Myth of Factive Verbs", *Philosophy and Phenomenological Research* 80: 497-522.
Heal, J. (2002). *Mind, Reason, and Imagination*. (Cambridge: Cambridge University Press)
Horgan, T. and Kriegel, U. 2007: "Phenomenal Epistemology: What is Consciousness that We Know it So Well?" *Philosophical Topics* 17: 123-44.
Hyman, J. 1999: "How Knowledge Works", *Philosophical Quarterly* 49: 433-51.
Lewis, D. 1995. 'Should a Materialist Believe in Qualia?' Australasian Journal of Philosophy, 73: 140-144, repr. in his Papers in Metaphysics and Epistemology (Cambridge: Cambridge University Press, 1999), 325-31.
McDowell, J. 1982: Criteria, Defeasibility, and Knowledge", *Proceedings of the British Academy* (68): 455-79. Also in J. Dancy, ed., Perceptual Knowledge, Oxford University Press, 1988.
----- 1995: "Knowledge and the Internal", *Philosophy and Phenomenological Research* (55): 877-93.
Moran, R. 2001: *Authority and Estrangement*. Cambridge: Harvard University Press.
Neta, R. and Rohrbaugh, G. 2004: "Luminosity and the Safety of Knowledge", *Pacific Philosophical Quarterly*
Neta, R. 2010: "The Nature and Reach of Priviliged Access", in Anthony Hatzimoysis (ed.) *Self Knowledge*. OUP.
Peacocke, C. (1996). "Entitlement, Self-Knowledge and Conceptual Redeployment." Proceedings of the Aristotelian Society **XCVI**: 117-158.
----- 1999: *Being Known*. Oxford: OUP.
----- 2008: *Truly Understood*. Oxford: OUP.
Pitt, D. 2004: "The Phenomenology of Cognition, or What is It Like to Think that P?" *Philosophy and Phenomenological Research* 69: 1-36.

Prinz, J. 2004: "The Fractionation of Introspection", *Journal of Consciousness Studies* (11): 40-57.
Pryor, J. 2005: "There is Immediate Justification" in M. Steup and E. Sosa, eds., *Contemporary Debates in Epistemology*, Blackwell.
----- forthcominga: "When Warrant Transmits", available at < http://www.jimpryor.net/research/papers/Wright.pdf>.
----- forthcomingb: "Uncertainty and Undermining", available at <www.jimpryor.net/research/papers/Uncertainty.pdf>.
Ryle, G. 1949: *The Concept of Mind*, London: Hutchinson.
Schiffer, Stephen. 2004: "Vagaries of Justified Belief," *Philosophical Studies*, 119: 161-84.
Schwitzgebel, E. forthcoming: "Acting Contrary to Our Professed Beliefs, or the Gulf Between Dispositional Belief and Occurrent Judgment"
Setiya, K. forthcoming: "Knowledge of Intention"
Shoemaker, S. 1995: 'Moore's Paradox and Self-Knowledge', Philosophical Studies, 77: 211 – 28
----- 1996: *The First-Person Perspective and Other Essays*. Cambridge: CUP.
----- 2009: "Self-Intimation and Second-Order Belief", *Erkenntnis* 71:35-51.
Siewart, C. 1998: *The Significance of Consciousness* Princeton University Press.
Silins, N. 2008: "Basic Justification and the Moorean Response to the Skeptic", *Oxford Studies in Epistemology*, vol. 2.
----- forthcominga: "Seeing through the 'Veil of Perception'", to appear in *Mind*.
----- forthcomingb: "Basic Self-Knowledge and the Problem of Easy Knowledge"
Sosa, E. 2003: "Privileged Access", in *Consciousness: New Philosophical Perspectives*, eds. Q. Smith and A. Jokic. Oxford: OUP.
Stoljar, D. 2009: "The Argument from Revelation" in *Conceptual Analysis and Philosophical Naturalism*. MIT.
Sutton, J. 2007: *Without Justification*. Cambridge: MIT.
Tye, M. 2002. "Representationalism and the Transparency of Experience," *Nous* 36, pp. 137-51.
Unger, P. 1975: *Ignorance.* Oxford University Press.
Vahid, H. 2005: "Moore's Paradox and Evans's Principle: A Reply to Williams", *Analysis*, 65: 337-41. Way, J. 2007: "Self-Knowledge and the Limits of Transparency", *Analysis* 295: 223-30.
Weatherson, B. 2008: "The Bayesian and the Dogmatist", *Proceedings of the Aristotelian Society*, 107 (2007): 169-85.
White, Roger. 2006: "Problems for Dogmatism", *Philosophical Studies*, 131: 525-557.
Williams, J. 2004: "Moore's Paradox, Evans's Principle, and Self-knowledge", *Analysis*, 64: 348-53.
Williamson, T. 2000: *Knowledge and its Limits*. Oxford: OUP.
----- 2004: "Skepticism", in F. Jackson and M. Smith (eds.) *The Oxford Companion to Analytical Philosophy*. Oxford: Oxford University Press.
Wright, C. 2002: "(Anti)-Sceptics Simple and Subtle: Moore and McDowell", *Philosophy and Phenomenological Research*, 65: 330-48.
Zimmermann, A. 2004: "Unnatural Access", *Philosophical Quarterly*, Vol. 54, No. 216 (July 2004), pp. 435-438.
----- 2006: "Basic Self-Knowledge: Answering Peacocke's Criticisms of Constitutivism", *Philosophical Studies*, 128 (March 2006) pp. 337-379.