

SANFORD C. GOLDBERG

THE PSYCHOLOGY AND EPISTEMOLOGY OF  
SELF-KNOWLEDGE

**ABSTRACT.** In this paper I argue, first, that the most influential (and perhaps only acceptable) account of the epistemology of self-knowledge, developed and defended at great length in Wright (1989b) and (1989c) (among other places), leaves unanswered a question about the *psychology* of self-knowledge; second, that without an answer to this question about the psychology of self-knowledge, the epistemic account cannot be considered acceptable; and third, that neither Wright's own answer, nor an interpretation-based answer (based on a proposal from Jacobsen (1997)), will suffice as an acceptable answer to the psychological question. My general ambition is thus to establish that more work is needed if we are to have a full account of self-knowledge in both its epistemological and psychological aspects. I conclude by suggesting how my thesis bears on those who aim to provide an empirical account of the cognition involved in self-knowledge.

1. INTRODUCTION

In the first section of this paper, I examine *the epistemology* of self-knowledge. I review Crispin Wright's case for the thesis that the epistemic justification of self-ascriptions of standing mental states derives from the fact that "by their very nature" mental states are "subject to groundless, authoritative self-ascription" (Wright 1989b, p. 630). Wright himself was aware that such an epistemically-minimalist view requires an account of the *psychology* of self-knowledge, concerning the manner in which we *come to form* first-person opinions; I conclude the first section by sketching Wright's own account on this score. In the second section, I examine the psychology of self-knowledge at greater length. I suggest that the 'self-regarding beliefs' that a person has – beliefs regarding the kind of person she is, the goals and aims she has, the moral standards to which she holds herself accountable – affect the first-person opinions she will be disposed to form; and I will gloss this by speaking of the 'cognitive penetrability' of first-person opinions. In the third section, I will return to Wright's account of the psychology of self-knowledge and will argue that it fails to square with the cognitive penetrability of first-person opinions. I will move on to consider an alternative (Wright-inspired) attempt to provide an account of



the psychology of self-knowledge, found in (Jacobsen 1997); I will suggest that this account, too, founders, though for a very different reason.

In my concluding section I will suggest that Wright's own account of the epistemology of self-knowledge is incomplete, as it has yet to be shown that this account can be made to square with what we know about the psychology of self-knowledge. I will speculate about the significance of the difficulty I am identifying; I do so in the hope that the speculations I offer will pave the way for the sort of further articulation I think we are owed by those who (like Wright) urge a minimalist approach to the epistemology of self-knowledge.<sup>1</sup>

## 2. THE EPISTEMOLOGY OF SELF-KNOWLEDGE

A person's opinions about her own standing states of mind seem to enjoy a unique epistemic status.<sup>2</sup> These opinions are highly reliable (i.e., are generally correct), and they appear particularly immune from (at least certain kinds of) error. Let us use the expression 'first-person opinions' to designate the relevant class of opinions, i.e., opinions that a person forms regarding her own standing states of mind. And let us speak of this unique epistemic status enjoyed by first-person opinions by speaking of 'first-person authority'.<sup>3</sup> On the assumption that there *is* something special about the epistemic status enjoyed by first-person opinions,<sup>4</sup> the epistemic question regarding self-knowledge is this: what accounts for first-person authority? I will begin by suggesting what sort of problem this question raises, and then I will suggest how Crispin Wright's own 'minimalist' conception of the epistemology involved appears to be uniquely suited to respond to this problem.

### 2.1. *The Problem of First-Person Authority*

In virtue of what do a person's opinions about her own standing attitudes (her first-person opinions) enjoy the secure epistemic status that they enjoy? What *makes* one's first-person opinions authoritative?<sup>5</sup> As (Wright 1989b and 1989c) make clear, this question turns into a *problem* once we appreciate that neither of the two traditional models of epistemic justification – the inference model and the observation model – can account for the authoritativeness of one's first-person opinions. Since these arguments are found in many other places,<sup>6</sup> I will only present as much of them, and in as crude a fashion, as serves my purpose of providing the basic motivation for Wright's minimalist account of the epistemology of self-knowledge.

Take first the claim that the model of inference will be unable to account for the phenomenon of first-person authority. According to the inference-based model, a particular belief counts as epistemically justified when (and to the degree that) the belief in question is the conclusion of some acceptable chain of reasoning or inference.<sup>7</sup> In brief, the problem with accounting for the epistemic status of first-person opinions on this model is this: one's judgements regarding one's own standing attitudes are often formed *without any evidence at all*, and so by extension *cannot* be represented as the conclusion of a process of acceptable reasoning from evidence. Now, it is certainly true that *sometimes* I may have to think hard about what I believe; and *sometimes* I may even have to collect evidence, typically in the form of my own behavior, in order to determine what I believe. The problem, of course, is that though these types of reasoning-based cases do exist, they are far from the standard way in which we form first-person opinions. On the contrary, in most cases I *just know* what I believe, without having to think about it or consider my recent behaviors. Someone asks me what I think about the New York Yankees' prospects for this year; without blinking an eye, I say that I believe that they will do very well. Or I am reporting my beliefs about the role of money in politics: I don't have to ponder at all to report that I believe that it's a shame how much influence money has in American politics.<sup>8</sup> In short, many – perhaps most – of a person's opinions about her own beliefs are not based on any evidence at all.<sup>9</sup> But if they are not *based* on any evidence at all, then they cannot be *inferred* from any evidence at all – in which case the inference model will be an inappropriate way to characterize their epistemic status.

Next, turn to the claim that (like the inference model) the observation model cannot account for first-person authority either. According to the observation model, a particular belief counts as epistemically justified, not by being an acceptable conclusion from a process of reasoning, but rather by being acquired via direct observation. One *just sees* that the sky is blue, *just hears* that the noise is loud.<sup>10</sup> Unfortunately, it is relatively clear that this model will not work for the judgements involved in self-knowledge, for the simple reason that there appears to be no *observable subject-matter* in self-knowledge. It is a now-long-discredited theory of mind which would treat our minds as private theaters, wherein we observe the goings-on with the mind's eye in a process known as 'introspection.' What is more, even if we still held to such a view regarding states of consciousness such as red afterimages and the like, it is not at all plausible that one's *beliefs* and *desires* are objects that parade through one's mental theater. Yet if it is not appropriate to treat beliefs as mental objects apprehended in some process akin to introspection, it is dubious at best whether we can even

*apply* the observational model to the realm of judgements concerning one's own beliefs. And so the observational model fails because there is nothing to *observe* in self-knowledge, in the way that there is a world to observe in perception.<sup>11</sup>

Having thus appreciated the difficulties facing anyone who would use either the inference model or the observation model to account for the epistemic status of first-person opinions, someone might be tempted by a simpler approach. The tempting idea is to say that first-person opinions are authoritative precisely because they are so *reliable*, and that *this is all that we can say on the matter*. Someone tempted in this direction would be urging a very thin account of first-person authority, one on which we take it as a *brute fact* that first-person opinions are reliable. Now, as we will see, there is something to be said for this position. However, at this point, taking the reliability of first-person opinions as a *brute fact* will not do. For, even if it is a *fact* that first-person opinions are reliable – and anyone who accepts the doctrine of first-person authority should agree that *this is a fact* – nonetheless taking this fact as *brute* will not let us address the natural question that arises: *why* are these opinions so reliable? What is there about this domain of facts, that a person's judgements about them are so reliable (normally true)? Since these question looks an awful lot like our original question – What makes first-person opinions authoritative? – the proposal to treat the reliability of first-person opinions as a brute fact is a refusal to face the issue, rather than an acceptable philosophical position.<sup>12</sup>

But now we have reached an impasse. We saw that giving an account of the epistemology of self-knowledge requires giving an account of the basis of first-person authority. But we face a problem insofar as the facts surrounding the formation of first-person opinions – the facts (i) that such opinions are typically formed without any evidence, and (ii) that they are not formed on the basis of observation – render the traditional models in epistemology incapable of accounting for the (by hypothesis, authoritative) epistemic status of first-person opinions. Since these are the only two traditional models of epistemic justification that we have, and since we cannot treat the reliability of first-person opinions as a brute fact, we appear to be without resources in our efforts to account for the authority of first-person opinions. What to do?

## 2.2. *Wright's Solution: The Avowability Conception of the Mental*

Following Wittgenstein's approach to matters of mind, Crispin Wright has advanced a proposal that is at once simple and radical. His proposal is based on the idea that the best way to approach these matters is by looking at the practices by which we *attribute* beliefs to ourselves and to others.

His thesis is that a careful look at what goes on when we attribute beliefs to ourselves and others, will suggest that the justification of a person's first-person opinions is *built into* the very concept of what it is to possess a mental state in the first place. Since this proposal looks a lot like the (recently rejected) proposal to treat the reliability of first-person opinions as a brute fact, it is worth examining the basis for Wright's thesis.

We have already noted two features of these practices. We have noted that, whatever the source of first-person authority turns out to be, first-person opinions are often *effortlessly formed* and typically *epistemically groundless*<sup>13</sup> (not based on any evidence at all). These points can be translated into the vocabulary of practice: it is part of our belief-attribution practices, that we treat as legitimate a person's self-attributions of belief even when these self-attributions are formed in an effortless manner and without any evidence.<sup>14</sup> To these two features of our practice of belief-attribution let us now add a third. The feature I have in mind derives from the possibility of self-deception. When do we treat a person as self-deceived? This question can be usefully framed in terms of evidence: what do we use as *evidence* of the falsity of a person's first-person opinions (i.e., of her being self-deceived)? While answers may vary, it is clear that the person's behavior, including her verbal behavior, provide the best (and perhaps the only) evidence we have for accusations of self-deception. So for example we might say: "Janet says that she isn't worried about the exam, but she gets visibly anxious when any mention is made of it, she has rearranged her schedule to leave lots of time to study for it, she spends all her free time at the library with books on the subject, etc." In short, we use behavioral evidence as the basis for accusations of self-deception; that we do so tells us that (it is a feature of our belief-attribution practices that) first-person opinions are *answerable*<sup>15</sup> to such evidence. In particular, when a person's particular first-person opinion fails to cohere well with her other behavior – what she has said and done in the past, and what she will say and do in the future – we can reach the verdict that the first-person opinion in question is *false* (i.e., that the person is self-deceived).

Here, then, we have three features first-person opinions: such opinions are typically *effortlessly made*, they are often *epistemically groundless*, and they are *answerable* to the person's own doings and sayings (past and future). Has this brief examination of what goes on in the attribution of belief gotten us any further in our attempt to account for first-person authority? On the contrary, it might seem as if we still have a problem on our hands, albeit a new one. The problem as we now see it is this. If the truth of a first-person opinion depends on its cohering with the sayings and doings of the person herself (as the point about the *answerability* of one's

first-person opinions to one's behavior would have it), then how can one's first-person opinions as a class be as *reliable* as they are? After all, we mentioned that the opinions themselves are made effortlessly and without any evidence; but then it would appear that they are (to put it vividly) *mere shots in the dark*;<sup>16</sup> and yet in order to be true they must cohere with the person's behavior in future circumstances – circumstances which the person herself typically has not had in mind as she formed her first-person opinion. It seems as if, given these points, first-person opinions should be very *unreliable*, not very reliable – and so it would seem that we are still far from the account we seek.

Wright himself has suggested that we can find a way out of our difficulty by shifting our perspective, from the practices surrounding the *self*-attribution of belief in first-person opinions, to the practices whereby we attribute beliefs *to others*. It is a central part of these practices, Wright notes, that all of us *do* assume that the people with whom we deal are authoritative about their own standing mental states. In fact, a close look at our belief-attribution practices suggests that it's not just that we assume that others are reliable *in general*; stronger still, we assume that what they are telling us regarding their own beliefs and desires is *true, unless we have determinate behavioral evidence to the contrary*. We give their word (in the form of their first-person opinions) *the presumption of truth*. In these terms, the epistemic question we now face is this: what justifies *us* in granting the presumption of truth to *other people's* first-person opinions, when those opinions are (as we have seen) (i) effortlessly formed, (ii) epistemically groundless, and (iii) answerable to the person's own subsequent sayings and doings – sayings and doings which the person herself rarely takes into consideration at the time she forms the first-person opinion itself?<sup>17</sup>

Wright's proposal is to treat the presumption of truth enjoyed by first-person opinions as part of *our very conception of what it is to possess a mental state*. The claim is that, such is the very concept of being in a mental state, that one's opinions about one's own mental states *stand by default* and are overturned only when we have determinate evidence to the contrary. If acceptable on other grounds, this theory would answer the reliability question: first-person opinions are as reliable as they are because they *stand by default*, only to be overturned in the case of determinate evidence to the contrary. In the very same way this theory would also provide the basis for the justification of the practice whereby we grant others' first-person opinions the presumption of truth.<sup>18</sup> In particular: others' first-person opinions are granted the presumption of truth as part of the very process by which we go about identifying others' intentional states in the first place.

Indeed, it is this latter claim – that the granting of authority to others’ first-person opinions is a part of the process whereby we identify the mental states of others – that prevents Wright’s theory from degenerating into a version of the idea that the reliability of first-person opinions is a *brute fact*. Wright himself has shown us how to see his proposal as something that need not be accepted as a *brute fact*, but rather as something that is motivated by a close look at the relation between the mental states themselves and the practices whereby we attribute mental states to ourselves and others.<sup>19</sup> His point, Wittgensteinian in its overtones, can best be brought out by contrasting one’s opinions about one’s own mental states, with one’s opinions about (say) the weather. Wright’s thesis is that one’s opinions about one’s own mental states (defeasibly) determine the facts about one’s mental states in a way that one’s opinions about (say) the weather do *not* (defeasibly) determine the facts about the weather.<sup>20</sup>

We can illustrate this contention as follows. Suppose that Alfred says, “It’s raining.” And suppose we want to know whether Alfred’s opinion is correct. Well, we don’t have to take *Alfred’s* word for it: we can just look outside the window. In this way we can determine whether the weather is as Alfred’s opinion would have it. But now suppose that what we’re investigating is not the weather, but Alfred’s own *beliefs about* the weather. And suppose Alfred says, “I believe that it’s raining,” and we want to know whether this opinion about what he believes is correct. Do we have any way to determine what he actually believes, independent of what he *tells* us on this score?

Perhaps it will be thought that we *do* have such a way. After all, above we suggested that nonverbal behavior is sometimes an indication of a person’s beliefs. So perhaps it will be thought that we can look to Alfred’s nonverbal behavior to determine what he believes. Now, to be sure, looking at his nonverbal behavior will give us a way in which to determine (at least for all practical purposes) which belief-attributions were *wrong*. So for example if he puts on his galoshes, chances are that it would be wrong to attribute to him the belief that it is sunny. But the problem is that a person’s nonverbal behavior is consistent with the attribution of *many different* belief’s and desires. For example, we can explain his putting on his galoshes in terms of:

- (a) the belief that it’s raining and the desire not to get wet;
- (b) the belief that his sister wants to see how his galoshes fit and the desire to please her;
- (c) the belief that his galoshes will soothe his aching feet and the desire for foot comfort; etc.

Now if we look at longer stretches of his behavior, we will be able to rule some of these out (again, at least for all practical purposes). But even after all the (non-verbal) behavioral evidence is in, there will still be many different belief-and-desire attributions that are compatible with his behavior. Indeed, virtually *any* belief-desire attribution can be made to cohere with – can be represented as rationalizing – a person’s behavior, so long as the attributor is willing to make suitable adjustments elsewhere in the corpus of beliefs and desires attributed to the subject under interpretation. This is just a special case of the doctrine of confirmation holism.<sup>21</sup>

We have just seen that nonverbal behavior alone will not tell us what in fact Alfred believes. Let us now show that *nothing short of taking Alfred at his word* will enable us to determine what he believes. Our problem was this: there are many different beliefs and desires which Alfred might be said to have, consistent with his nonverbal behavior. So how to determine what he *actually* believes and desires? How are we to determine which, of the various belief-desire attributions consistent with his nonverbal behavior, is the correct one? The obvious answer, which is probably the *only available* answer, seems to be this: *we must take his word on it*. Except in cases of determinate evidence to the contrary (in the form of a failure of his avowals<sup>22</sup> to square with his behavior), whatever Alfred *says* he believes is what he *does* believe. *Short of making this assumption there will be no way to determine which, of the many belief-desire attributions consistent with his nonverbal behavior, is the correct one.*

Notice that this point about how we determine the facts of Alfred’s psychology stands in stark contrast to our attempt to determine the facts of the weather. There, we saw that we need not take Alfred’s word for it – we could just check the weather ourselves. But when the facts in question concern Alfred’s psychology, *we must treat his opinions about these facts as presumptively true*, on pain of *not being able to determine these facts at all*.

Let us summarize these points, in an attempt to characterize what we have learned about the epistemology of self-knowledge. The epistemic task was to account for first-person authority, i.e., the authoritative epistemic status enjoyed by first-person opinions. We saw that neither of the traditional models of epistemology will give us the desired account of this status. Then we saw that we *can* have the desired account if we accept the thesis that by their very nature first-person opinions are presumptively true, overturned only when there is determinate (behavioral) evidence to the contrary. Finally, following Wright’s lead, we saw that this thesis is not a version of the reliability-as-brute-fact idea, but rather is supported by an examination of how one’s opinions about one’s beliefs are related

to the facts about what one believes. The epistemology of self-knowledge that emerged from this examination was a *minimalist* conception, according to which the authority of first-person opinions is *built into* our very conception of what it is to have a mental state. This epistemology is ‘minimalist’ in the sense that the epistemic warrant for first-person opinions is not seen as deriving from some kind of ‘truth-tracking’ manner in which first-person judgments are formed. Rather, *what the facts themselves are* is itself defeasibly determined by *the judgments* that the person is disposed to form about these facts: one’s first-person opinions are held to stand by default, overturned only in cases where there is determinate behavioral evidence to the contrary.<sup>23</sup>

### 2.3. *One Remaining Problem: Explaining the Fact of Coherence*

It is easy to see, however, that this account of the epistemology of self-knowledge is not complete as it stands. In particular, we are left with one remaining question. We just said that, while all first-person opinions enjoy the *presumption* of truth, not all are ultimately *true*: in particular, those that cannot be made to cohere with one’s behavior are overturned. In line with this we can now express the conclusion of the previous section in terms of the following two conditional claims (holding for any person *S* and any first person opinion to the effect that one oneself believes that *p*):

If *S*’s first-person opinion that she herself believes that *p* *does* cohere with *S*’s sayings and doings, then that first-person opinion determines the relevant facts (i.e., *S* is treated as *in fact* believing what she *says* she believes);

if on the other hand a particular first-person opinion of *S*’s *does not* cohere with her sayings and doings, then this first-person opinion is overturned (i.e., *S* is treated as *not* in fact believing what she *says* she believes, that is, *S* is treated as self-deceived).

And now it is clear that the epistemic account from the previous section cannot be the full story. For we still want to know *why it is* that in point of fact one’s opinions *do* manage to cohere so well with one’s sayings and doings in circumstances so far unconsidered.<sup>24</sup> This striking fact – let us call it the ‘fact of coherence’ – is something that has yet to be explained. Yet the account on offer needs to explain the fact of coherence; failure to do so would make it seem miraculous that our first-person opinions manage to cohere with our behaviors, in future situations not considered at the time at which the first-person opinions were formed. In short, failure to answer this question about the fact of coherence is tantamount to leaving the epistemic account offered so far to hinge on something miraculous.

Wright’s own explanation of the fact of coherence is rather quick (1989b, 632–3). He believes – largely (I suspect) because he thinks that

his theory *forces* him to believe – that there must be some ‘subcognitive’ mechanism that is responsible for the production of our first-person opinions, and that it is this mechanism that is responsible for the fact that our first-person opinions manage to cohere so well with our future (verbal and nonverbal) behaviors. The proposal is that first-person opinions are like opinions about (for example) one’s own body-position. As these opinions are understood by many contemporary cognitive scientists, one’s body-position opinions are not arrived at by way of reasoning from evidence, but rather our brains have mechanisms that (on receiving physical inputs from our muscles and other bodily parts) simply produce the opinions we do have regarding the present position of our body.<sup>25</sup> We assume that there is an evolutionary story to be told as to why we have such mechanisms. But in any case we are ‘forced’ in a sense to have those body-position opinions that we do have – their formation is not something that is under our rational control. Wright’s thesis is that first-person opinions are in the same category: in the basic cases these opinions are the product, not of any rational deliberation on our part, but rather a brain mechanism that is responsible for this domain of opinion-formation.

So stated this explanation for the fact of coherence needs to be filled in. But I shall not criticize this proposed explanation as being incomplete; rather, I think that it is unacceptable *in principle*. I say this in full awareness that Wright’s proposed explanation is an empirical hypothesis, one whose truth is to be determined by empirical investigation and not by *a priori* philosophical reasons. It is just that I think that the empirical evidence that we *do* have, together with the theory we bring to bear in thinking about this evidence, together suggest that Wright’s proposed explanation (even in its sketchy form) *cannot be the full account of the fact of coherence, even if the proposed explanation were to be filled out completely*. This is a strong claim. In order to justify it, I will have to move from considerations concerning the *epistemology* of self-knowledge, to considerations concerning the *psychology* of self-knowledge.

### 3. THE PSYCHOLOGY OF SELF-KNOWLEDGE

First, a story. Consider the case of Ed. Ed was raised in a deeply religious household. He was raised to believe in God, and to think of himself as one of God’s noble creatures. Before he came to college, Ed was seen by many of his classmates as a bit too earnest and self-righteous. They complained, for example, that he would not acknowledge anything but the most innocent and morally pure motivations; that he would not confess to believing anything but what the Bible taught him; and that he treated

Scripture as the literal word of God. Then Ed went to college, where he studied Philosophy and Psychology. He thought that this would reinforce his whole religion-centered belief system. He was wrong.

As luck would have it, he mistakenly signed up for a class in Freudian psychology. Too lazy to switch out of the class, Ed remained in the course. Curiously, rather than being repulsed or outraged, he began to find himself strangely attracted to Freud's theory. At first his attraction was more a matter of idle curiosity – Ed was intrigued by the strange and curious views heathens could hold. But as time went on idle curiosity gave way to suspension of judgement, and finally he began to entertain seriously the possibility that Freud had it right after all. In this process, he began to change his views about himself. He began to wonder whether all of his motivations were as innocent as he had previously thought. He had begun to see himself capable of all sorts of sexual desires to which he would never before admit. And so on.

What can we make of Ed's case, regarding *the psychology* of self-knowledge? One thing seems clear. Given the radical nature of the change in his self-conception, Ed will now avow having beliefs and desires which before he would have denied; and, conversely, Ed will now deny having certain beliefs and desires which before he would have avowed. So for example, now when the topic turns to the subject of his mother, Ed says such things as "I believe that she's smothering me with all sorts of sexual attention," "I wish that Father were not so severe in his dealings with me when I am near Mother," and so on. Before coming to college, he would never have said such things – on the contrary, he would have been outraged by the attribution.

We can formalize this point a bit as follows. I will use the term 'self-regarding belief' to refer to any belief about one's own moral and psychological constitution. Included among these beliefs are beliefs about the sort of person one is, about the aims and ambitions one oneself has, about the moral standards to which one holds oneself responsible, and so on. My thesis about the psychology of self-knowledge, then, is that *certain changes in one's self-regarding beliefs will affect the first-person opinions one forms (is disposed to form)*. To be sure, real life sees very few cases of the sort of radical changes as were exhibited in the example of Ed. But there are more realistic, if less vivid, examples of this same phenomenon. Suppose for example that Vivian changes her self-image: while she used to think of herself as not among the most intelligent person in her class, she now thinks that this view of herself was limiting her effectiveness in class, and so now has a more optimistic self-appraisal. Then upon being queried about a particular math puzzle that was announced in her math class, she

says, “I believe that it’s a very interesting puzzle” – whereas before she might have said, “I think it’s silly.” Once again, we have illustrated the same point: changes in certain of one’s self-regarding beliefs affect what first-person opinions one is disposed to form.<sup>26</sup>

None of these claims is all that controversial. At least I hope they aren’t. I now want to argue that these claims, uncontroversial as they are, are incompatible with Wright’s proposed explanation of the fact of coherence. The fact that these claims are not controversial will thus forestall any thought about preserving Wright’s answer by denying the claims themselves.

#### 4. PUTTING THE EPISTEMOLOGY AND THE PSYCHOLOGY TOGETHER

Let’s return to Wright’s proposed explanation of the fact of coherence. Why is it that, for any given person *S*, *S*’s first-person opinions generally cohere with her future doings and sayings in situations so far unconsidered by her (i.e., at the time of forming those opinions)? Wright’s explanation was that there is a *subcognitive mechanism* that is responsible for the formation of these opinions. Unfortunately, when we turn to the question how to distinguish between mechanisms and processes that are cognitive from those that operate at a level below that of cognition,<sup>27</sup> we see that *the manner in which the cognitive/subcognitive distinction has been drawn is incompatible with the claim that there is a subcognitive mechanism responsible for first-person opinions*. In particular, I will now argue that, given what we have already said about the psychology of self-knowledge, and given the almost universally-accepted criterion for the postulation of a subcognitive mechanism, the fact of coherence cannot be exhaustively explained by appeal to a subcognitive mechanism.

Let us begin by characterizing the almost universally-accepted criterion for the cognitive/subcognitive distinction. What makes a belief-forming mechanism subcognitive? The most widely-accepted answer to this question, defended at length in Pylyshyn (1984) is that a belief-forming mechanism is subcognitive only if the mechanism is “cognitively impenetrable.”<sup>28</sup> A mechanism is *cognitively impenetrable* (and so subcognitive) only if the mental states that the mechanism produces as its output *cannot be influenced by any of the background beliefs that the agent herself has*. As Pylyshyn puts the point, a process can be said to be *cognitively impenetrable* (and so subcognitive) only when its behavior “requires no explanation in terms of semantic regularities – that is, in terms of rules and representations” (Pylyshyn 1984, 130–1) – and he contrasts such processes with those processes that exhibit “rationally explicable alterability [in their

output] *in response to changes in goals and beliefs*" (Pylyshyn 1984, 133; italics mine). I will first explain this characterization of cognitive impenetrability, and then I will show that the criterion of cognitive impenetrability is violated in the case of first-person opinions.

Return for a moment to the class of first-person opinions regarding one's own body-position. Presumably, the opinions that one forms regarding the position of one's own body are not affected by the other beliefs one has: change a person's background beliefs as much as you like, that person's first-person opinions regarding the position of her own body will be invariant through these changes.<sup>29</sup> Now, it is an empirical question whether there are any belief-forming mechanisms that answer to the description of being cognitively impenetrable. Many people in cognitive science think that there are such mechanisms. The best example is most probably the mechanisms involved in early vision (Pylyshyn 1984, 135). But, whether or not these *are* examples of subcognitive mechanisms, the point remains that if *there are to be* such subcognitive mechanisms, then (prevailing opinion is that) such mechanisms *must be cognitively impenetrable*. That is, such a mechanism must work in such a way that the beliefs that it produces as its output<sup>30</sup> are invariant over changes in one's other beliefs; such a mechanism must produce these beliefs in a manner that is *insensitive* to the person's other beliefs.<sup>31</sup>

The thesis that we have been considering, then, is that the cognitive impenetrability of the process by which we form a certain class of judgements is a necessary condition<sup>32</sup> for the existence of a subcognitive mechanism responsible for the production of that class of judgements. Suppose we accept this claim. Well, it is pretty clear that this criterion for the postulation of a subcognitive mechanism is violated in the case where the opinions in question are first-person opinions about one's own mental states. Recall our thesis regarding the psychology of self-knowledge: it was that *certain changes in one's self-regarding beliefs will affect the opinions one forms about one's own beliefs, desires, and so on*. And recall that Pylyshyn himself characterizes cognitive penetrability in terms of the "rationally explicable alterability of a [process's] behavior *in response to changes in goals and beliefs*" (again, italics mine). In light of this characterization of what it is for a belief-forming process to be cognitively penetrable, it is clear that my thesis about the psychology of self-knowledge amounts to the thesis that the process by which we form first-person opinions is *cognitively penetrable*. But now we see that, if the process by which we form first-person opinions *is* cognitively penetrable, then the fact that we form the first-person opinions that we do cannot be explained – at least cannot be exhaustively explained<sup>33</sup> – by appeal to a subcognitive mechanism.

Now, one might think to avoid my conclusion asserting the cognitive penetrability of the processes by which we form first-person opinions, by presenting an alternative characterization of what it is for a process to be cognitively penetrable.<sup>34</sup> One might think to construe cognitive penetrability as follows: a belief-forming process is cognitively penetrable only if *a specification of its input fixes its output*. This characterization captures the idea that, such are cognitively impenetrable (subcognitive) processes, that they can be seen as functions from input to output: once their input is specified, nothing (and so no further changes in background belief) will affect their output. The potential benefit of this characterization of cognitive penetrability is clear: on such a characterization, changes in background belief *can* affect the output of the process by which we form first-person opinions, without thereby undermining the claim that such processes are cognitively impenetrable, for the simple reason that those background beliefs are themselves *part of the very input* of the process by which we form first-person opinions. And so this alternative conception of cognitive impenetrability could be used to undermine my central claim; for my central claim has been that the processes involved in first-person opinion-formation are cognitively penetrable *because* changes in self-regarding beliefs affect the output of such processes.

But this alternative conception of cognitive penetrability is not acceptable. This much is clear when we remind ourselves of the context in which Pylyshyn himself motivated the concept of cognitive impenetrability in the first place. This concept was formulated so as to capture the distinction between processes that are cognitive from those that are subcognitive. Further, the difference between cognitive and subcognitive processes is conceived by Pylyshyn and many others to be a difference between two distinct levels of explanation. While the subcognitive level involves explanations that are either biological or else merely syntactic (Pylyshyn 1984, 131), at the cognitive level explanations are *rational* explanations, i.e., explanations that advert beliefs and desires and their rational relations to behavior. It is for this reason that Pylyshyn himself characterizes cognitive penetrability, which is supposed to indicate *cognitive* (as opposed to subcognitive) processes, in terms of the “*rationally explicable alterability* of a [process’s] behavior *in response to changes in goals and beliefs*.” But notice that the way in which Pylyshyn has connected the cognitive/subcognitive distinction with the idea of different levels of explanation (rational vs. nonrational) is precisely what gets lost on the alternative conception of cognitive penetrability. For, on the alternative conception, it would be *false* that subcognitive processes do not exhibit the “*rationally explicable alterability*” of which Pylyshyn speaks; on the contrary, they

would exhibit such alterability in precisely the cases involving first-person-opinion-formation in the face of changes in one's self-regarding beliefs. Since this implication cuts against the very point behind the concept of cognitive penetrability, and since it appears that the only justification for the alternative conception (which has this implication) is to be able to resist the conclusion I have drawn, it would appear that the alternative account itself is unwarranted. In short, the burden of proof would be on those who seek to resist my conclusion asserting the cognitive penetrability of the processes responsible for first-person-opinion-formation; they must show that there is an alternative conception of cognitive penetrability available, one which does not cut against the motivation for the concept of cognitive penetrability in the first place.

Before concluding, however, I should identify and reject a more radical reaction to my argument so far. The preceding reaction was to accept the cognitive penetrability criterion for subcognitive processes but simply to reject Pylyshyn's conception of that criterion. The more radical reaction is to reject the cognitive impenetrability criterion in the first place. If such a reaction were justified, then we could continue to insist on Wright's own proposed explanation for the fact of coherence (i.e., the appeal to a suitably-rigged subcognitive mechanism) in the face of evidence for the cognitive penetrability of first-person opinions. Nonetheless I think that this option is a desperate and *ad hoc* move, one that no principled proponent of a cognitivist theory of mind ought to endorse. To be sure, the criterion in question embodies what is after all a theoretical claim; namely, the claim that all subcognitive mechanisms are cognitively impenetrable. As a *theoretical* claim this claim might be false, and there is no simple, direct way to test it. But I would pause to suggest that it appears to be a claim which most working cognitive scientists accept, or *ought* to accept at any rate (see Pylyshyn 1984, Chapter 5). So even though the cognitive penetrability criterion *might* be an unacceptable criterion, it would be unwise for someone to reject this proposed criterion unless she had good and independent reasons for thinking that it *is* unacceptable. Surely the desire to save Wright's approach to the epistemology of self-knowledge is not reason enough. In short, given the fruitfulness of the cognitivist research program of which this criterion is a central component, the move to reject this criterion is *ad hoc*.

To conclude, it would appear that the cognitive penetrability criterion for subcognitive processes must be accepted: and that, short of having an acceptable alternative conception of cognitive penetrability in hand, Pylyshyn's conception of cognitive penetrability stands. Yet we have seen that, given Pylyshyn's conception of cognitive penetrability, the processes

responsible for the formation of first-person opinions *are* cognitively penetrable. The result, of course, is that Wright's own explanation of the fact of coherence (postulating a subcognitive mechanism responsible for first-person-opinion-formation) is unacceptable *because incompatible with the psychology of self-knowledge*. But I indicated above that if Wright has no explanation of the fact of coherence, then the epistemic account he has provided cannot be considered a complete account of the epistemology of self-knowledge: in such circumstances his account would be based on the unexplained fact of coherence. Given that Wright's theory is arguably the only acceptable account of the epistemology of self-knowledge, we thus find ourselves facing a difficulty: our best (and perhaps only) account of *the epistemology* of self-knowledge raises a question about how speakers manage to form first-person opinions that cohere with their future behavior, in advance of their considering the circumstances of those behaviors; and yet Wright's proposed explanation for this fact of coherence is incompatible with what we can reasonably take to be *the psychology* of self-knowledge.

##### 5. AN ALTERNATIVE INTERPRETATION-BASED EXPLANATION FOR THE FACT OF COHERENCE: JACOBSEN 1997

I now want to turn to one reaction to this difficulty, found in Jacobsen (1997), which is worth dwelling on in some detail. My claim shall be that this proposal is actually based on an unacceptable circularity, as emerges when we consider how such an account would explain the fact that one's avowals cohere as well as they do with one's previously-expressed opinions and avowals.

So far, the general difficulty I have raised for Wright's minimalist epistemology of self-knowledge derives from a question about the manner in which we come to form first-person opinions. The question concerns how we manage to form these opinions in such a way that they largely cohere with our future doings and sayings, despite the fact that we typically form these opinions without giving the circumstances of those future doings and sayings any thought. My claim was that Wright's answer to this question is incompatible with the fact that first-person opinions are formed in a cognitively penetrable fashion. But perhaps we can leave Wright's epistemology in place and come up with an *alternative explanation* for the fact of coherence. Now, so far as I can tell, I know of only *one alternative explanation* which has ever been seriously put forward by those who have followed this strategy. The alternative explanation is to see the coherence in question as largely (if not entirely) *imposed from without* in the process of interpretation, and so not something that has to be explained by postulating

a suitably-rigged subcognitive system. Such a view is a natural part of Davidson's theory of radical interpretation.<sup>35</sup> Perhaps more controversially, it can be said that John McDowell too subscribes to the interpretation-based explanation: in his 1991 criticism of Wright's explanation for the fact of coherence, McDowell appears to suggest that the coherence in question is imposed as the cost of treating another as intelligible to a scheme of interpretation *at all*.<sup>36</sup> But perhaps the most explicit endorsement of the interpretation-based account of the fact of coherence is found in Jacobsen (1997). In an article otherwise sympathetic to Wright, Jacobsen (pp. 436–9) explicitly proffers an interpretation-based explanation for the fact of coherence, and goes so far as to suggest that such an explanation actually emerges from Wright's own work.<sup>37</sup>

Let us begin, then, with the interpretation-based explanation itself. To understand this account we would do well to review what it is that we expect to be explained. As philosophers interested in self-knowledge we begin by noting that, as a matter of fact, people's first-person opinions cohere remarkably well with their future behavior, and that this coherence obtains even though in most cases people form their first-person opinions without giving any thought at all to the circumstances of their own future behavior. For this reason we suppose that this 'fact of coherence' reflects a *remarkable fit* between one's first-person opinions and one's future behavior.<sup>38</sup> It is this remarkable fit that we would like to have explained. We think to ourselves: how can this be, that a person's first-person opinions manage to cohere so well and so often with her future behaviors (including verbal behaviors), even though she herself typically does not give any thought to those behaviors or the circumstances in which she exhibits them? Indeed, we can now see why Wright himself felt the need to postulate a subcognitive mechanism involved in the production of first-person opinions. He postulated such a mechanism precisely *because* he acknowledged that the fit between first-person opinions and subsequent behavior is remarkable indeed, believing as he apparently did that only an appropriately-rigged subcognitive mechanism could explain this (otherwise-remarkable) fit.

But one aspect of the line of reasoning that leads to the postulation of a subcognitive mechanism ought to be highlighted. The fit between one's first-person opinions and one's behavior is 'remarkable' only on the assumption that a person's behavior establishes substantial constraints on what will count as cohering with this behavior. The alternative explanation for the fact of coherence is based on the idea that behavior does *not* in fact exert substantial constraints on what counts as cohering with that behavior. The idea is that the fit (between first-person opinion and future

behavior) is unremarkable to *begin with*, not because there is a suitably-rigged subcognitive mechanism at play, but rather because the coherence itself is the byproduct of the process whereby we interpret other speakers. To see this, we need only remind us of the kinds of consideration that had motivated Wright's minimalist epistemology in the first place. He had noted that one's non-verbal behavior alone will not determine what one believes (i.e., many different and incompatible belief-attributions will square with one's behavior); and he had urged that it is a person's own first-person opinions that play the constitutive role in (defeasibly) determining, among the set of belief-attributions that square with one's behavior, which of those attributions is in fact the correct description of what one believes.

Given these two points we get the clear sense that Wright's appeal to a subcognitive mechanism was in fact superfluous. Indeed, the impression we now have is that the so-called 'remarkable fit' is not actually remarkable at all, simply because *first-person opinions have a constitutive role in the determination of the facts of a person's psychology*. For it now seems as though the fact of coherence is nothing more than a reflection of the fact that, when we interpret another, we come to the scene *forced* to treat the person's first-person opinions as presumptively true. And this point has the effect that we come to the scene *forced to make the behavior cohere with these opinions as much as possible* – forced, that is, if we hope to be able to identify what it is that an agent actually believes and desires.<sup>39</sup> One point here is worth underscoring. The coherence in question is an *artifact of interpretation*, something that is *imposed by interpreters* during the course of interpretation. As such, the fact of coherence is not something that *the person herself* has to achieve, nor is this coherence something that some *subcognitive component* of her mind has to be rigged to achieve for her.

There is an alternative way to present the proposed interpretation-based explanation for the fact of coherence. In a general sense, a person's first-person opinions *have to cohere* with her behavior, *on pain of not being recognized as first-person opinions at all*. It isn't as if there is a creature who might have all sorts of wild first-personal opinions – first-person opinions that fly in the face of *its* future behaviors – thereby showing us that we humans do something remarkable in having first-personal opinions that manage to cohere (or 'fit in') with *our* future behaviors. Rather, if there were a creature whose supposed "first-personal opinions" were so outlandish as to fail to be able to cohere at all with its behavior, then we would conclude that in fact the creature was not reporting first-personal opinions *at all*.<sup>40</sup> In this manner we see that the "coherence" in question is largely an imposed coherence. With this point in place, the fact of coherence seems

less remarkable, and so less in need of the sort of explanation Wright has offered.

Now, many people would simply stop at this point, leaving the impression that the fact of coherence is exhaustively explained as an artifact of the methodology involved in interpretation and belief-attribution. We see this most clearly in Jacobsen (1994) and less clearly in McDowell (1991; see footnote 36); but my guess is that those sympathetic to Davidson's project in Radical Interpretation would concur that the above account pretty much says (at least in outline form) all that can be said about the fact of coherence. Against them, I aim to suggest that this alternative account actually explains less than what its proponents suppose: even if what it says is true, this explanation does not provide – indeed, given its resources, may in fact be incompatible with – a full explanation of the fact of coherence. After suggesting why this is so, I will argue that this point ought to temper the enthusiasm with which proponents of the interpretation-based account endorse the account itself; at the very least they owe us more explanation if their account is not to be left presupposing a miracle.

### 5.1. *Against the Interpretation-Based Account of the Fact of Coherence*

Our dialectic so far has been this. In order to be accepted as true, the attribution to a person of a mental state must cohere with her future behavior (verbal and non-verbal). In this general sense, then, a person's future behavior constrains the mental-state attributions that are true of that person. But above we saw that, at this level, the behavioral constraint on acceptable attributions is not very substantial: while one's behavior might rule out *some* candidate belief-attributions as incorrect, *many other* candidate belief-attributions will be consistent with that behavior. Thus we came to suppose that the behavioral constraint on acceptable attributions, while real enough, is not all that substantial, and that it is one's self-attributions of belief (one's first-person opinions) that play the central role in *determining* which, of the many candidate belief-attributions that cohere with one's behaviors, is correct. These points led us to think that we should treat the fact of coherence as largely if not entirely imposed from without, an artifact of the process of interpretation.

But this picture of things makes the fact of coherence seem *less remarkable* than it actually is. We can bring this out by restricting our focus to that aspect of the fact of coherence whereby one's present first-person opinions generally manage to cohere with one's past and future *utterances*. Clearly, this point is general: for *any* arbitrary person *S*, *S*'s first-person opinions somehow manage to cohere largely with her

(*S*'s) own previously-expressed opinions (including previously expressed *first-person* opinions).

Now consider the number of garden-variety examples we could come up with to illustrate just this point. On virtually any topic on which *S* actually has held some attitude or other, if someone were at the present time to query *S* concerning *S*'s attitudes about the topic, in most cases *S* would simply "know" these attitudes without giving them any thought. Further, *S* could at present express her attitudes in a way that coheres with the vast body of her previously-expressed opinion on the subject, again, without giving those previously-expressed opinions any thought. Let me now suggest how this kind of consideration, humdrum and uncontroversial as it is, can be used to show that the coherence-as-artifact-of-interpretation view appears to be unacceptable. My claim will be that this account has no non-circular explanation for this aspect of the fact of coherence.<sup>41</sup>

What explains the fact that our present first-person opinions cohere as well as they do with our previously-expressed opinions (including our first-person opinions)? Surely this fact cannot be explained away as an artifact of interpretation. The methodology of interpretation only tells us that *if* we are to be interpretable at all, then there must be some general sense that can be made of our vocalizations; this methodology does *not* tell us what explains why *in point of fact* it *always* turns out that some general sense can be made of our vocalizations. So for all that interpretation explains, we could have been creatures that made noises from time to time, such that the noises we made over any interval simply were not systematic enough to be interpreted as speech acts at all, let alone evaluated on grounds of how well these speech acts cohered with one another. The fact that we are most decidedly not such creatures – that our utterances *can* be interpreted as speech acts, and that, among other things, we avow standing attitudes in a way that coheres remarkably well with our previous avowals and expressions of opinion – still requires an explanation

I now want to argue that the most natural explanation of this fact – which arguably is the only *available* explanation – is not open to a proponent of the interpretation-based approach to the propositional attitudes. The natural explanation is this. On the assumption that a thinker's beliefs do not change over time, then what it is that enables the thinker to avow all and only those beliefs which cohere with her previous avowals and expressions of opinion (again, allowing for occasional self-deception or insincerity) is nothing other than *the having of the (avowed) beliefs themselves*. It is *because* *S* continues to believe that *p*, that she is presently disposed to *avowing* the belief that *p*. On such a proposal, sameness of belief across time explains sameness of avowal dispositions across time.

This explanation seems well and good. Indeed, it is unclear whether there even is another way to explain the fact that (for all thinkers at all times) a thinker's present dispositions to avow are dispositions to avow beliefs which largely cohere with her previously-expressed opinions and avowals.

But notice that this explanation undermines the central insight (if it is indeed an insight) of the interpretation-based account. For above we saw that, in attempting to explain the original fact of coherence, the interpretation-based account was led to treat dispositions to avow as (defeasibly) determining the facts about what a thinker believes. Unfortunately, with such an analysis in place, it is nothing other than a *vicious circle* to cite the fact that *S* continues to believe that *p* as part of the explanation for why she has the disposition to avow precisely the belief that *p* (rather than one or another of the many belief-contents that would not cohere with her past opinions and avowals). In effect, doing so would be to treat the truth-conditions for '*S* believes that *p*' as defeasibly determined by *S*'s avowal-dispositions, only to go on to explain why *S* has the avowal-dispositions she does by appeal to *S*'s having the beliefs that she does! Surely such a circle is vicious.

At this point it is perhaps worth reminding ourselves why the interpretation-based account is committed to the problematic analysis in question. When it was first presented this analysis was presented as an insight, i.e., as the only way to account for the authoritativeness of first-person opinions. The idea was that it is only by treating one's first-person opinions as themselves defeasible determinants of what one believes, that one could manage to account for the authoritativeness of those opinions in the face of a failure to account for such authoritativeness using any of the traditional models of epistemic justification. However, we now see that such an analysis comes at some cost, since if it is one's avowal-dispositions that defeasibly determine what one believes, then one cannot cite one's continuous having of the beliefs one does in an explanation of how one's avowals manage to cohere so well with one's previous avowals and previously-expressed opinions. The circularity charge looms.

Now what might a proponent of the interpretation-based account say in response? While I cannot claim to know all possible reactions, I can imagine three main ones. I will argue that none appears promising.

The first reaction would be to suggest that, while it is correct that the proponent of the interpretation-based account cannot cite the having of beliefs in an explanation of this aspect of the fact of coherence, nonetheless this is no problem, since something *other* than the having of beliefs explains (the problematic aspect of) this fact. This may be so, but we are owed some explanation of what this something is. In short, this first reac-

tion is essentially an empty promissory note until some concrete proposal is offered.

The second reaction would be to *accept* that it is the persistent having of beliefs that explains this aspect of the fact of coherence, while arguing that, despite appearances, the circle I have identified is not vicious. But I cannot see how this reaction is to be rendered plausible. If to believe that *p* just is to be disposed to *avow* believing that *p* (so long as one's avowal coheres with one's behaviors), as the interpretation-based account would have it, then I cannot see what there can be to *believing that p* which might explain the having of the disposition to avow such a belief. In short, this reaction would appear to be a non-starter.

The third (and, on the face of it, most promising) reaction would be to try to reconstruct the interpretation-based account so that it is compatible with a more robust conception of what it is to believe that *p*, such that on this more robust conception *believing that p* might plausibly be used to explain *having the disposition to avow believing that p*. Here is an occasion to return to an idea mentioned in a footnote above, involving the conception of beliefs as sentences-in-a-belief-box. While this is not the only possible way to attempt to pursue the third reaction, it is worth looking at: first, because the sentences-in-a-belief-box (or SBB) conception of belief is widely thought to be plausible on independent grounds;<sup>42</sup> second, because on the face of it such a marriage looks promising; and third, because its failure (or what I will argue is its failure) is instructive of what I take to be the problems facing *any* attempt to follow the third reaction itself.

To begin, the SBB-conception of belief appears to provide the proponent of the interpretation-based account precisely what she needs: a conception of beliefs on which beliefs have a reality that is independent of the believer's avowal-dispositions, such that beliefs *so conceived* might plausibly be thought to explain those dispositions. The idea is this. To have the belief that *p* is to have a token-sentence  $S_1$  in one's belief box, where  $S_1$  means that *p*. Now, having the token-sentence  $S_1$  in one's belief-box has certain effects. Central among these effects is the effect that, under certain further conditions (i.e., thinking about or being asked what one believes), one will have in one's belief box another sentence  $S_2$ , where  $S_2$  itself *contains* a token-sentence of the same type as  $S_1$ . In short, on this model, first-person opinions are reliable (assuming one does not change the opinions in question) precisely because among the typical effects of having a belief is that one will believe that one has that belief: this is spelled out by saying that it is among  $S_1$ 's typical effects that  $S_1$ 's being in the belief box typically causes  $S_2$ 's being in the belief box.

Unfortunately, in the present context, the SBB conception of belief will not help with the problem that faces the interpretation-based account of the fact of coherence: for either the SBB conception of belief itself is a *competitor* with the interpretation-based account, and so cannot be used to supplement that account without calling into question the very motivation for that account; or else if the SBB conception of belief is modified so as to leave some work for the interpretation-based account, then the resulting proposal can be seen to face the same problem as the one I indicated above.

Take the first horn. If beliefs are taken to be sentences in a belief-box, and if the reliability of first-person opinions is traced to the fact (or what this conception *alleges* is a fact) that among the typical effects of having a given sentence in one's belief box is that one will typically acquire a higher-order belief to the effect that one has that first-order belief, then we have a right to ask: why then do we need the interpretation-based account at all? What work is being done by the central contribution of that account, namely, the idea that one's avowal dispositions defeasibly determine what one believes? Isn't it the case that, on the contrary, what one believes is determined by the sentences that are in one's belief box? In short, on one reading of the proposal to supplement the interpretation-based account of the fact of coherence with the SBB conception of beliefs, the interpretation-based account itself is an extra wheel.<sup>43</sup>

But now there does appear to be a way to marry these two proposals, while leaving work for the interpretation-based account. The idea would be this. The sentences that figure in one's belief-box can be individuated in one of (at least) two ways. They can be individuated syntactically, i.e. without reference to their content; or they can be individuated semantically, with reference to their content. The former kind of individuation ensures that beliefs are real (and so have causal powers) independent of how we determine their content; and this is sufficient to enable us to appeal to a particular belief to explain the disposition to avow having that belief. At the same time, we are not in a position to identify any particular belief as the belief that *p* (for example), until such time as we have arrived at the proper interpretation of the sentences themselves – something for which we will have to avail ourselves of whatever method we bring to bear in interpreting the thinker's public utterances. This ensures that there is a role for the interpretation-based account to play, with the result that the first horn of the dilemma is avoided.<sup>44</sup>

But now I think that we face the second horn of the dilemma, which is that we appear to be facing precisely the same problem that the appeal to the SBB conception of belief was supposed to enable us to avoid. For consider. Plausibly, it is only the having of the belief *that p* that could serve as

a candidate explanation for the having of the disposition to avow believing that *p*. (Not just any old syntactically-individuated belief will do.) But now what makes something the belief that *p*, as opposed to a belief with some other content, is precisely that a proper theory of interpretation, as applied by way of the thinker's public utterances to the sentences that figure in her belief box, tells us that a mentalese sentence of this particular syntactic structure means that *p*.<sup>45</sup> And if this is true then it seems as though we are back to the problem we were hoping to avoid: for once again we appear to be explaining why one has the disposition to avow believing that *p* by appeal to the having of the belief that *p*, where the having of the belief that *p* (as opposed to some other belief) itself is individuated in part by appeal to the having of certain linguistic dispositions.

Notice that this particular problem which the second horn presents is not one that would face a theorist not already committed to the interpretation-based account. For a proponent of the SBB conception of belief might plausibly maintain that what explains the coherence of the beliefs one is presently disposed to avow (on the one hand) with one's previously-expressed opinions and avowals (on the other) is this: the very same syntactic structure, that both (i) is the vehicle of one's first-order belief and (ii) causes one to have the higher-order belief regarding one's first-order belief, is *simply inherited* by the syntactic structure that is the vehicle of one's higher-order belief. Insisting on the idea that one's second-order belief inherits part of its syntactic structure from the very first-order belief which it purports to specify would ensure that one's first-order belief, *whatever its content*, will be correctly specified by the second-order belief.<sup>46</sup> Unfortunately for the proponent of the interpretation-based account, however, such a proposal seems to leave that account with no work to do (as in the first horn of the dilemma); all of the work in explaining the fact of coherence would be done by beliefs as syntactically (rather than semantically) individuated. Thus it would appear that someone who is antecedently committed to the interpretation-based account cannot simultaneously avoid both horns of the dilemma.

In sum. None of the three ways in which a proponent of the interpretation-based account might think to react to my charge of circularity – accept the circularity but deny that the circularity itself is relevant to the explanation, deny that there is even an apparent circularity, or else show how the apparent circularity is merely apparent (by borrowing a more robust notion of belief) – appears plausible at present. With regard to the last of the three reactions, I have only shown how *one* particular way to try to beat the wrap (i.e., by appeal to the SBB-conception of belief) does not look promising; I have not shown that *no* way to beat

the wrap will be forthcoming. But since I for one cannot think of another way to do so, what I have shown should at the very least convince that the burden of proof lies with the proponent of the interpretation-based account. What is more, it should convince that what such a proponent must do is no easy task: she must formulate a notion of belief substantial enough to explain a thinker's avowal dispositions, but insubstantial enough not to undermine the analysis to which the interpretation-based account is committed. Absent a proposal that satisfies these demanding (if not simultaneously-impossible-to-satisfy) constraints, what I have shown here, I think, is that the interpretation-based account is without a non-circular account of the fact of coherence. I believe that this is a serious criticism of the interpretation-based account, since without a non-circular explanation for the fact of coherence this account is left making that fact seem miraculous.<sup>47</sup> If true, this would establish that the interpretation-based account does not solve the problem that prompted Wright to appeal (ultimately unsuccessfully) to a subcognitive mechanism in the first place.

#### 6. A PROGRAMMATIC CONCLUSION

In this paper I have sought to identify a difficulty confronting those who would seek to offer a full theory of self-knowledge in both its epistemic and psychological dimensions. The basis of the difficulty is this. Our best account of the epistemology of self-knowledge, based on the idea that first-person opinions are groundlessly formed, tempts us to treat the facts about a person's beliefs as themselves (defeasibly) determined by the person's avowal-dispositions. But such an account raises a question concerning the coherence of one's avowals (first-person opinions) with one's previous and subsequent verbal and non-verbal behavior. The problem arises when we see that, given certain facts about the psychology of self-knowledge, there would appear to be no explaining (what otherwise seems miraculous) this fact of coherence.

One aspect of this problem (that treated in Section 4) is strictly philosophical, since it has to do with the task of rendering compatible the commitments of the interpretation-based account of the fact of coherence with the requirements on a notion of belief if beliefs are to be part of the explanation for the fact of coherence. But another aspect of this problem (that treated in Section 3) is broadly empirical, insofar as it has to do with the way in which to theorize about the psychology involved in self-knowledge. In an effort to speak to those who would attempt to model the cognition involved in self-knowledge, I want to conclude by reflecting on the significance of the latter aspect of the problem.

To my mind, one striking theme emerges from the argument of Section 3: what the epistemology of self-knowledge gives with one hand (i.e., the idea that the justification of our first-person opinions is groundless, *because the formation of these opinions does not implicate our other beliefs*), the psychology of self-knowledge appears to take with the other (i.e., the formation of our first-person opinions *does* implicate our other beliefs, as reflected in the fact that the processes responsible for first-person-opinion-formation are cognitively penetrable).

At one level, the way out of this difficulty is clear. It is that one's first-person opinions should be seen as *informed by*, though *not justified on the basis of*, one's other beliefs. Indeed, I submit this as the main positive claim of this article: first-person opinions are formed in a manner that is (*epistemically*) *groundless yet cognitively penetrable*. But this turns out only to be a superficial resolution of the difficulty, merely *putting a label* on it rather than *resolving* it. For it is the groundless-yet-cognitively-penetrable aspect of the process whereby we form first-person opinions that makes it difficult to account simultaneously for the epistemology and the psychology of self-knowledge. This is because, having acknowledged the category of judgement into which I have been assimilating first-person opinions, we must then acknowledge that the process of belief-formation can draw on one's other beliefs in ways other than by having those other beliefs as *premises* from which the belief formed is *consciously or unconsciously inferred*.

What difficulties remain, once we acknowledge the distinction between the *epistemic groundlessness* of first-person opinions and the *cognitive penetrability* of the process whereby we form these opinions? I believe that the remaining difficulties, whatever they are, are essentially of a piece with the difficulties involved in coming up with an adequate empirical theory that models the cognition involved in self-knowledge. In this respect my thesis here has the status of a constraint on what is to count as an acceptable model. If an acceptable model is to be had, it must make clear how the epistemic groundlessness of an opinion is compatible with the idea that the process of forming the opinion is sensitive to the thinker's stock of beliefs. If a proposed model cannot make sense of this idea, then it cannot claim to be a correct model of the cognition involved in self-knowledge: for it will not have succeeded in getting its account of the psychology of self-knowledge to square with an acceptable account of the epistemology of self-knowledge. And yet to make sense of the idea that first-person opinions are epistemically groundless yet cognitively penetrable, the model must confront the question I have identified: how *can* a process of opinion-formation work in such a way that one's other beliefs inform the opinion

that is yielded by the process, but *not* in a way that implicates those other beliefs in the epistemic appraisal of the opinion yielded?

It would go some distance towards convincing us that this state of affairs is indeed possible, if it should turn out that first-person opinions are not the *only* type of opinions whose formation is epistemically-groundless-yet-cognitively-penetrable. Perhaps we can look to one's first-person *memory* reports as having these same features.<sup>48</sup> But establishing this parallelism, and drawing the right lessons from it, are topics I must reserve for another occasion. Instead, I will conclude with a comment on the importance of carrying out the task I have identified. Prior to having a clear account of the psychology of self-knowledge, we cannot even be sure that Wright's minimalist epistemology of self-knowledge *can* be made to square with the psychology of self-knowledge. And this point, together with the failure of two distinct attempts to show how Wright's account *could* be made to so square, should temper our enthusiasm for what may be the only acceptable account of the involved epistemology. It is in the context of this looming paradox that I suggest that more work needs to be done.<sup>49</sup>

#### NOTES

<sup>1</sup> In addition to Wright 1989b and 1989c, proponents of a minimalist epistemology of self-knowledge include Davidson 1988; McDowell 1991; and Shoemaker 1994.

<sup>2</sup> More formally, the opinions in question are those expressed by present-tense self-ascriptive judgements of the form 'I believe [worry; fear; hope] that *p*.' To be sure, a fully general account of the epistemology of self-knowledge would have to cover one's knowledge of one's own *occurrent thought* and *sensory states* as well, in addition to one's knowledge of one's standing attitudes. But I will restrict myself to knowledge of one's own standing attitudes, out of the belief that the judgments involved in the latter forms of knowledge possess features unique to them, and so must be treated separately. (Regarding the special features of the judgments involved in self-knowledge of *occurrent thoughts*, see Goldberg 1997a, Goldberg (1999), and Goldberg (forthcoming).)

<sup>3</sup> I employ this deliberately vague expression out of an interest to avoid controversy: saying that first-person opinions are 'authoritative' is meant to be compatible with whatever is one's preferred story regarding their epistemic status.

<sup>4</sup> Not everyone subscribes to this assumption, but I will not defend the assumption here. First, I think that the vast majority of philosophers *would* indeed subscribe to it; and second, the assumption has some intuitive plausibility in its own right. These two points suggest that the assumption should be accepted as part of our working theory, to be jettisoned only if there are good (independent) reasons for doing so. With this in mind the present paper is aimed at those who would accept the assumption as a good starting point for theorizing.

<sup>5</sup> Throughout this paper (unless otherwise noted), for the sake of simplicity I will restrict my attention to those thinkers whose opinions have not changed over time; for any such change in opinion will require a similar change in *first-person* opinion (if the thinker's first-

person opinion is to count as true), thereby needlessly complicating our analysis. I thank an anonymous reviewer for indicating to me the need to make this comment.

<sup>6</sup> These arguments are presented with great subtlety in Wright's own work (see his 1989a; 1989b; and 1989c); Shoemaker 1994 (see especially Lectures I and II); Burge 1996; and it is plausible that arguments for this view are also presented in Davidson 1987 and McDowell 1991.

<sup>7</sup> The chain itself need not be conscious in one straightforward sense of 'conscious': if I acquire the belief that it will snow by listening to a radio forecast of snow, my belief that it will snow is justified by the lights of the inferential model even if I did not consciously think to myself: "The reporter on the radio said it will snow; radio reporters are typically reliable; I have no reason to think that this reporter was being anything but sincere; so I believe that it will snow."

<sup>8</sup> To be sure, *memory issues* may be involved in cases in which I have previously come to believe something and at the present time report it upon being queried on the subject. But this does not affect the crucial point: even if memory issues are involved, nonetheless when I report my first-person opinion now, that opinion is not formed on the basis of any kind of *inference* – at least not if 'inference' is used in the sense relevant to the inference-model in epistemology (where the validity of the inference is what grounds the justification of the belief so inferred). (I thank Andrew Pessin for stressing to me the importance of memory in first-person opinion reporting.)

<sup>9</sup> In my concluding section I will more directly characterize the notion of a belief's being 'based on evidence.' But I will not do so here, for a reason that touches on the aim of this paper. It is my view that the problem of first-person authority presents us with a particular kind of difficulty when we try to formulate this notion with precision; and it is because I will try to *argue* for this view that I postpone characterizing the notion itself. For now I am relying on an intuitive but merely implicit understanding of the notion.

<sup>10</sup> This model can be made more complicated in various well-known ways. For instance, we might say that what makes these examples of seeing and hearing examples of *epistemically justified belief*, is that the beliefs we form on the basis of our seeing and hearing are formed by way of a *reliable sensory apparatus*. The criticism that follows in the text is meant to apply to *any* version of the observation model.

<sup>11</sup> The forgoing paragraph is no doubt crude in the extreme, and there are many subtle and interesting things that can be said on the issue whether it is appropriate to think of self-knowledge on the model of observation. But it is not my goal to rehearse the moves and countermoves that can be made in this regard; I refer those interested in this topic to Wright 1989b and 1989c, the first two lectures in Shoemaker 1994, and Burge 1996, pp. 108–110.

<sup>12</sup> There is one other way in which some philosophers might think to account for the reliability of first-person opinions, which will have occurred to those who are attracted to the beliefs-as-sentences-in-a-belief-box model. I will wait until Section 4 to examine such a proposal at length. (I thank Andrew Pessin and Adam Vinuesa, both of whom, independently of one another, suggested such an account to me in conversation.)

<sup>13</sup> Throughout I will drop the qualifier 'epistemically'. My use of the term 'groundless' is borrowed from Crispin Wright 1989b.

<sup>14</sup> Let me make explicit what this talk of practice comes to. To say that these constitute part of our belief attribution *practices* is to make a comment about the sorts of thing that would pass for a legitimate move in our game of belief-attribution. So, to stick with our example, to say that it is part of our belief-attribution practices that the self-attribution of

belief is effortless, is to say that it is a legitimate move in the game of belief-attribution to express a first-person opinion *prior* to having expended any effort to investigate into the matter what one oneself believes – and, as a corollary, that in typical situations it would *not* be a legitimate move in our game of belief attribution to challenge *someone else's* first-person opinion merely because that person had formed her first-person opinion prior to having performed any investigation into the matter of what she believes. (Having provided the translation to the vocabulary of practice, however, I will refrain in what follows from speaking in this vocabulary – I find the vocabulary of practice rather unnatural – but it is to be understood that what I say can be translated into the vocabulary of practice in the general way I've indicated here.)

<sup>15</sup> This use of the term 'answerable' is borrowed from Crispin Wright 1989c.

<sup>16</sup> It might be thought that our first-person opinions are *not* 'shots in the dark,' so long as the beliefs we report in these opinions are beliefs we *recall* having previously formed. This may be so. But consider first-person opinions that report beliefs which the thinker has *not* previously formed. If we restrict our attention to such opinions, these *will* have all of the outward features of being 'shots in the dark.' Since I trust that there will be *many* first-person opinions that fall into this category, the description seems apt. (I thank Andrew Pessin for pointing out the need for this comment.)

Note too that it will not do to describe such cases (where one is avowing a belief that was never previously formed) as the person's *deciding* what she believes, and leave it at that. For if the person in question "decides" that she believes that *p*, and yet behaves in such a way that attributing to her the belief that *p* does not cohere with the rest of her behavior, then we would treat her as being *wrong* about what she believes. In short: even if we treat cases of never-before-formulated belief as cases of *decisions*, we can still ask: but do these decisions square with the person's future behaviors? This alone ought to convince us that treating first-person opinions as "decisions" does not prevent them from having the character of 'mere shots in the dark.'

<sup>17</sup> This formulation of the problem is Wright's; see his 1989c.

<sup>18</sup> In fact, as Wright himself points out in his 1989b and 1989c, this practice of granting the presumption of truth to other's first-person opinions is not merely justified; it is justified *a priori*. It is not as if we have to wait to see how well a person's first-person opinions cohere with her future doings and sayings, *prior* to treating these opinions as presumptively true; on the contrary, we treat her opinions as presumptively true right from the start. While this point does figure in the manner in which Wright motivates his avowability conception of mental states, it need not detain us further here.

<sup>19</sup> Nor is the argument to follow the *only* argument that Wright gives for his avowability conception of the mental. He also argues that no other conceptions of mental states will be able to explain why it is that we are *a priori* justified in granting the presumption of truth to others' first-person opinions; see Wright 1989b and 1989c, as well as Chapter 6 of Goldberg and Pessin 1997 for my review of this argument.

<sup>20</sup> This very thesis is echoed in the first two lectures of Shoemaker 1994.

<sup>21</sup> To be sure, the principle of simplicity would tell against the more fanciful attribution schemes, favoring as it does the simpler attribution schemes over more complicated ones. Nonetheless, we should not feel comfortable resting the decision between attribution schemes on the principle of simplicity alone, for various reasons. First, while it is true that *in general* simpler schemes are to be preferred to more complex ones, nonetheless *in particular cases* there may be grounds for preferring more complex schemes. Second, there may well be cases involving two distinct schemes of (roughly) equivalent simplicity – in

which case the principle of simplicity will not help us decide which of the two is the correct one. Finally, over-reliance on the principle of simplicity, as what determines the correctness of an attribution-scheme, will raise worries on the score of realism regarding the entities being attributed (beliefs and desires). In short, even acknowledging the important role that the principle of simplicity plays in helping us decide between alternative attribution schemes, our point still stands: in at least some cases, mere (non-verbal) behavior together with the principle of simplicity will not succeed in singling out what by intuitive standards is the correct attribution scheme.

<sup>22</sup> I should be clear that, while I will use the term ‘avowal’, I do so as a stand-in for the longer ‘first-person opinion’; in doing so I do *not* mean to be subscribing to the view, championed by those who consider themselves ‘avowal theorists’, that first-person present-tense uses of psychological verbs have only an *expressive* (and not a *descriptive*) function.

<sup>23</sup> For more on this notion of the dependence of *the facts* regarding one’s standing psychological states on one’s *judgements* regarding these facts, see Wright 1989c and Shoemaker 1994.

<sup>24</sup> For noting the importance of this question, as well as for formulating the question in this manner, I am indebted to Wright himself. For the former see his 1989b, p. 632. For the latter see his 1989c Section 4.

<sup>25</sup> Or so we are told by some physiologists (McCloskey 1978), cognitive neuroscientists (Roll, Roll, and Velay 1991) and cognitive psychologists (the various contributions in Berinudez, Marcel, and Eilan 1995). The literature on proprioception (as this type of transduction is known) is huge; many other recent sources can be found in the references provided by the work I’ve cited.

<sup>26</sup> I would only add that contemporary psychological studies attest to our thesis. While the literature is vast, much of the recent literature is summarized in a review article, Sedikides and Skowronski (1995). I quote from this article:

... [T]he cognitive representation of the kind of person we think we are ... can ... influence the ways people perceive and remember other people ... ; lead people to defend themselves against threatening events and ideas ... ; [and to] determine future plans ... or other future-oriented thought ... (p. 245).

Clearly, much of this confirms our thesis that certain changes in one’s self-regarding beliefs will affect the opinions one forms about one’s own beliefs, desires, and so on. For consider. If one’s ‘self-representation’ – Sedikides’ and Skowronski’s term for what I am calling one’s set of ‘self-regarding beliefs’ – can influence the ways we perceive and remember other people, then presumably one’s self-representation can influence the beliefs we take ourselves to have regarding other people. (A parallel point can be made regarding future-oriented thought.)

<sup>27</sup> For my purposes here I will be following Pylyshyn’s understanding of the distinction between cognitive and subcognitive, as the difference between two distinct levels of explanation. Loosely, a state of the mind/brain is *cognitive* if and only if it is a state whose causal role in the intentional behavior of the organism requires to be explained in terms of the state’s *semantic content*; while a state of the mind/brain state is sub-cognitive if and only if it is a state that has a role in the production of intentional behavior, yet whose role in such production can be explained without reference to the semantic content of the state itself. This characterization of the distinction is slightly different from tradition, which had it that the cognitive/subcognitive distinction is the distinction between states of the mind-brain that are semantically-evaluable and those that are not. My reason for preferring Pylyshyn’s conception of the distinction is that (it seems to me that) his conception can

acknowledge a truth that the traditional conception cannot acknowledge: the truth, namely, that a brain state can be semantically-evaluable without it being the case that it has its role in the production of behavior in virtue of its semantic content. However, I have no axe to grind here; so readers who prefer the traditional conception of the distinction can read that conception into my argument. The important point is that the cognitive/subcognitive distinction is central to the very concept of the “functional architecture” of the cognitive system as a whole and the concept of functional architecture is a core part of contemporary theory in cognitive science (for which see Pylyshyn 1984, 259–62).

<sup>28</sup> Fodor (1983) uses the term ‘informationally encapsulated’ to make the same point.

<sup>29</sup> This is not to say that *nothing* can affect one’s first-person body-position opinions. Under certain extremely stressful conditions (including but not limited to the heat of battle during war), a person may not immediately acknowledge changes in his own bodily position, in a way that he *would* immediately acknowledge these changes under more normal conditions. Stories are told, for example, of men who do not realize that they have lost a limb until after the heat of the battle has passed; and often they do not realize that the loss of limb has occurred until it has been called to their attention by others. I thank Dr. Kathleen Rockland of the University of Iowa’s Department of Neuroscience for pointing this out to me.

<sup>30</sup> It would be more correct to speak of the ‘mental representations’ which it produces as its output; belief is *an attitude taken towards* such representations, and so is beyond what is produced by the mechanism itself. Since my point above could be easily modified in this way, for stylistic purposes I will continue to speak of ‘beliefs’, ‘judgements’, and ‘opinions’ rather than ‘mental representations’ as the output of the mechanism.

<sup>31</sup> It might help matters if we present a case in which the belief-forming mechanism clearly is not cognitively impenetrable. Consider for example the formation of opinions about the figure skaters on the US figure skating team. Suppose that Jones is asked to state his opinions concerning who is the best skater on the team. Suppose Jones tells us that *X* is the best skater around. Now suppose that Jones is told that *X* has lost two of her last three competitions in which she has participated to *Y*, another skater on the same team. Finally, suppose that Jones comes to believe what he is told. Would acquiring the belief that *Y* has beaten *X* in two of the last three head-to-head competitions affect Jones’ opinion about who is the best skater? Quite possibly. What this shows us is that whatever it is that is responsible for the formation of figure-skater-opinions clearly *is* sensitive to the background beliefs Jones has. We describe this by saying that the formation of opinions about figure skaters involves a process that is *cognitively penetrable*. And the cognitive penetrability of the process whereby one forms one’s figure-skating opinions would be taken by cognitive scientists as confirming the hypothesis that there is no subcognitive mechanism responsible for figure-skater-opinions (Indeed, who would have thought that there was such a mechanism?)

<sup>32</sup> It may well be a *sufficient* condition as well; but since this involves taking a stronger stand, and since taking the stronger stand is not required by my argument, I will be neutral on this score.

<sup>33</sup> I want to allow for the possibility that subcognitive mechanisms do play *some* role in the formation of first-person opinions. My point here is to insist merely that such mechanisms, by themselves, cannot be *responsible for* our forming the opinions we form. An alternative way to put this same point: citing such mechanisms could not constitute a *sufficient explanation* of our forming the first-person opinions we form.

<sup>34</sup> This objection is owed to Andrew Pessin.

<sup>35</sup> By this I do not mean that *Davidson himself* endorses, or would endorse, the view that the fact of coherence is imposed from without, in the process of interpretation. I suspect that *he would* endorse such a view; but even if he wouldn't his theory can be seen as providing the basis for doing so.

<sup>36</sup> I say 'arguably,' since it is unclear whether McDowell is committed only to the *general* thesis that too much incoherence in a purport system of beliefs entails that such a system is not a system of beliefs after all, or whether he is committed in addition to the *stronger* (because more specific) thesis that, given the methodology of interpretation, there can be no further questions about how a thinker forms a coherent self-conception once we see the role of charity in interpretation. As will emerge in this section, I think that the first (general) thesis is correct but that the second (stronger, more specific) thesis is false. And so, if all McDowell is committed to is the general thesis, then he will not be a target of the argument of this section.

<sup>37</sup> I should admit that I myself have been somewhat optimistic about the prospects for the interpretation-based explanation for the fact of coherence (see Goldberg and Pessin 1997, Chapter 6, Section 5). I now think that this view is unacceptable as it stands, for reasons I am about to consider.

<sup>38</sup> The *fit* is seen in the coherence of the opinions with the behavior, and the *remarkableness* of this fit is seen in that this coherence obtains despite the fact that people do not stop to consider the future circumstances of their behavior.

<sup>39</sup> The point here is that the fit in question is imposed in the process of identifying what it is that the subject believes and desires: failing to impose this fit will result in failing to be able to select the correct attribution to make from among the many belief-desire attributions which square with the agent's behavior (as discussed above). This point is developed and defended at length in Jacobsen (1997, 436-9), who suggests that the point itself was precisely what Wright himself claimed, when he claimed that mental states by their very nature are subject to groundless, authoritative self-ascription. In light of this Jacobsen suggests that Wright, having made just this point, had at his disposal the proper explanation for the fact of coherence, but for some reason failed to exploit it.

<sup>40</sup> For a similar suggestion, see McDowell (1991, 166-7) and my (Goldberg 1997b).

<sup>41</sup> While a previous version of this paper had suggested that the interpretation-based account is without a plausible explanation of the fact of coherence, I thank an anonymous reviewer of this journal for criticizing my argument in that previous version, and for suggesting the manner in which I should try to bring this point out in the present version.

<sup>42</sup> In saying this, of course, I am not endorsing the SBB conception. (For what it's worth, I'm skeptical.)

<sup>43</sup> It might be wondered: if the belief-as-sentence-in-belief-box idea appears as promising as it does as an explanation for first-person authority, then perhaps *this* is the kind of answer we should be looking for as our answer to the problem of squaring the epistemology of self-knowledge with the psychology of self-knowledge. Actually, it's not obvious that this would be a good answer, for two reasons.

The first reason, ironically enough, is that such an account would appear to work *too well*, i.e., it would appear to make a mystery of motivated *self-deception*. If all that it takes to obtain self-knowledge is for one's first-order belief that *p* to cause a higher-order belief that one oneself believes that *p*, and if the latter kind of belief is represented as a distinct sentence in the belief box which itself contains as a proper part a sentence type-identical to the sentence expressing the first-order belief, then we need to know what would account for self-deception. For consider: sometimes we have the belief that *p* but (for whatever

reasons) fail to have the higher-order belief (alternatively: fail to have the disposition to avow having this belief). How could a conception of beliefs as sentences-in-a- belief-box, as supplemented with the idea that a typical effect of a sentence's being in the box is that one acquires the higher-order belief to the effect that one has that first-order belief, accommodate the fact of self-deception? This question is not meant to be rhetorical, but rather is meant to point out what such a conception of first-person authority would owe us. It would owe us an account of how it is determined, given all of our beliefs, which beliefs are such that they give rise to the having of the correct second-order belief, and which are not – and why (i.e., under what circumstances). That this will not be easy is seen in the fact that not *every* belief is a likely candidate for self-deception, but rather only those that are somehow psychologically important to us. (Indeed, this very consideration might be taken to suggest part of the attraction of the interpretation-based account.) Perhaps such an account can be formulated, but no such account at present exists. (See Goldberg 1997b for an argument for a more skeptical conclusion towards the possibility of such an approach to self-knowledge and first-person authority, developed on the basis of this kind of consideration.)

The second reason why it is not clear that the SBB-conception of first-person authority will prove acceptable is that the job of explaining the fact of coherence goes beyond merely showing why we are disposed to avow having the beliefs we actually have. In particular, we must also explain why it is that we typically are *not* disposed to avow having beliefs whose contents are such as to fail to cohere with the contents of the beliefs we *do* have. For example, someone who believes that squash is disgusting and that squash is responsible for all of her present neuroses, typically will not (without some complex rationale) avow the belief that squash production should be increased. This, too, needs to be explained, since this aspect of coherence, too, is achieved without giving any thought to one's past avowals and previously- expressed opinions. No doubt, part of the explanation will be that in general we do not *have* beliefs whose contents fails to mesh with the contents of the beliefs we do have. But citing this as an explanation seems to me to beg the question, since we might reasonably wonder why we typically do not have such (content- anomalous) beliefs. Failing to explain this feature of our mental lives would appear to leave the SBB-conception of first-person authority with presupposing something miraculous in *its* turn. Only in this case, the miracle would be that as a matter of contingent fact, we generally do not have beliefs that fail to cohere with one another. Notice that, on the interpretation-based view, this 'fact' was not contingent at all: on the contrary, on such a view, what explains this 'fact' is that, if there were a creature whose "beliefs" were such as to be below some threshold of coherence, then we would conclude, *on the basis of our methodology of interpretation*, that such a creature is not a creature with beliefs at all. It would seem, then, that here is the strength of the interpretation-based account, vis-a-vis the SBB-conception's account of first-person authority. To summarize this second point: somehow our avowal-dispositions are content-sensitive, and this cannot be straightforwardly explained merely by appeal to the syntactic approach outlined above. So it is not clear that the SBB-conception's approach to first-person authority will amount to a full explanation of what we want explained. (I hope to be able to expand on this topic in another paper.)

<sup>44</sup> Actually, much of Fodor's "Substitution Arguments and the Individuation of Beliefs" (Fodor 1990, Chapter 6) appears to suggest that Fodor himself might be sympathetic to this kind of position; see especially the last paragraph on p. 172. However, I cannot claim to be certain of this attribution.

<sup>45</sup> I should point out that Fodor himself, in the article mentioned in the footnote above, explicitly rejects this account as an account of the individuation of mental content. (He

explicitly distinguishes between the individuation of belief-states and the individuation of mental contents, and holds (1) that the interpretationist picture only offers a theory of the former, and (2) that nothing much regarding the latter follows from this.) I note this only to suggest that Fodor's position on this score is *so much the worse for the proponent of the interpretation-based account*; for his position only succeeds in showing that the first horn of the dilemma facing such an account is not so easily avoided.

<sup>46</sup> Strictly speaking, I should include the following modification: *so long as syntactic structure plus context determine content*. I bring this up because of worries that might arise given a commitment to an externalism about mental content. In any case, I should add that this very conception of the relation between one's first- and second-order intentional states is itself part of a strategy that Burge and others have famously used to try to argue for a compatibilism between the doctrines of mental content externalism and authoritative self-knowledge of content. For criticisms and limitations of these kinds of attempts see my 1997a, 1999, (forthcoming).

Perhaps I should remind the reader as well that the result mentioned in the main text (i.e., that a proponent of the SBB-conception who is not also a proponent of the interpretation-based account need not fear the second horn of the dilemma) does not clinch the SBB conception's account of self-knowledge. As I mentioned in a footnote above, there are other problems that face such an account. All that I have shown here is that those problems are not identical to the problems facing the interpretation-based theorist.

<sup>47</sup> I thank an anonymous referee for suggesting that I formulate the criticism in this manner.

<sup>48</sup> I owe this suggestion to Andrew Pessin.

<sup>49</sup> I would like to thank the participants of various conferences at which I have given earlier versions of this paper (the 1995 Iowa Philosophical Society Conference, the 1996 Mid-South Philosophy Conference, the 1996 Southern Society for Philosophy and Psychology Conference, and the 1998 Central States Conference); the faculty and students of the Philosophy Department of Kenyon College, in front of whose Philosophy Club I gave a portion of this paper; Dr. Kathleen Rockland and Dr. Ralph Adolphs of the Department of Neuroscience at the University of Iowa, for helpful discussions of the nature of the problem I am trying to identify; various philosophers and psychologists, for the help they gave me in thinking through some of these issues (including but not limited to Ray Elugardo, Janet Gibson, Eric Kraemer, Ulf Nilsson, Bill Robinson, Robert Stainton, Gene Witmer, and Adam Vinueza); my good friend Andy Pessin, who deserves a special thanks for comments that were both quite extensive and extremely insightful; and finally, two anonymous referees for this journal, for their very helpful criticisms and suggestions.

#### REFERENCES

- Bermudez: 1995, *The Body and the Self*, MIT Press, Cambridge, UK.  
 Burge, T.: 1996, 'Our Entitlement to Self-Knowledge', *Proceedings of the Aristotelian Society* **96**.  
 Fodor, J.: 1983, *The Modularity of Mind*, MIT Press, Cambridge, UK.  
 Fodor, J.: 1990, *A Theory of Content and Other Essays*, MIT Press, Cambridge, UK.  
 Goldberg, S.: 1997, 'Self-Ascription, Self-Knowledge, and the Memory Argument', *Analysis* **57**, 3.  
 Goldberg, S.: 1997b, 'The Very Idea of Computer Self-Knowledge and Self-Deception', *Minds and Machines* **7**, 4.

- Goldberg, S.: (1999), 'The Relevance of Discriminatory Knowledge of Content', *Pacific Philosophical Quarterly* **80**, 2.
- Goldberg, S.: (forthcoming), 'Externalism and Authoritative Knowledge of Content: A New Incompatibilist Strategy', *Philosophical Studies*.
- Goldberg, S. and Pessin, A.: 1997, *Gray Matters: An Introduction to the Philosophy of Mind*, (M.E. Sharpe, Armonk, NY).
- Jacobsen, R.: 1997, 'Self-Quotation and Self-Knowledge', *Synthese* **110**, 419–445.
- McCloskey: 1978, 'Kinesthetic Sensibility', *Physiological Review* **58**, 783.
- McDowell: 1991, 'Intentionality and Interiority in Wittgenstein', in Puhl (ed.), *Meaning Skepticism*, de Gruyter, New York.
- Pylyshyn, Z.: 1984, *Computation and Cognition*, MIT Press, Cambridge, UK.
- Roll, Roll and Velay: 1991, 'Proprioception as a Link Between Body Space and Extra-Personal Space', in Paillard, Jacques et al. (eds.), *Brain and Space*, Oxford University Press, Oxford.
- Sedikides and Skowronski: 1995, 'On the Sources of Self-Knowledge: The Perceived Primacy of Self-Reflection', *Journal of Social and Clinical Psychology* **14**, 3.
- Shoemaker, S.: 1994, 'Self-Knowledge and 'Inner Sense': Lectures I–III', *Philosophy and Phenomenological Research* **54**, 2.
- Wright, C.: 1989a, 'Critical Notice: Wittgenstein on Meaning', *Mind* **48**, 390.
- Wright, C.: 1989b, 'Wittgenstein's Later Philosophy of Mind: Sensation, Privacy, and Intention', *Journal of Philosophy* **89**, 622–34.
- Wright, C.: 1989c, 'Wittgenstein's Rule-following Considerations and the Central Project of Theoretical Linguistics', in George (ed.), *Reflections on Chomsky*, Blackwell, Oxford, UK.

Department of Philosophy  
1427 Paterson Office Tower  
University of Kentucky  
Lexington, KY 40506-0027  
E-mail: scgold@pop.uky.edu

