Experience, Appearance, and Hidden Features

D. Gene Witmer
Department of Philosophy
P.O. Box 118545
University of Florida
Gainesville, FL 32611-8545
U.S.A.

gwitmer@phil.ufl.edu

PSYCHE, 7(09), July 2001
http://psyche.cs.monash.edu.au/v7/psyche-7-09-witmer.html

KEYWORDS: Philosophy of mind, consciousness, a priori, reference, context, analysis, concepts, zombies.

COMMENTARY ON: C. Siewert (1998) *The Significance of Consciousness.* Princeton: Princeton University Press. x + 374pp. ISBN: 0-691-02724-2. Price: $42.50 hardcover.

ABSTRACT: Charles Siewert has given us an ingenious thought experiment involving a limited lack of conscious experience. The possibility of the described case is incompatible with a number of popular theories of consciousness. Siewert acknowledges, however, that this possibility is not a direct threat to "hidden feature" theories. I aim to do two things: first, strengthen his defense of the claim that the case is genuinely possible by considering and rejecting some further attempts to explain away our temptation to believe it possible; and second, to explore how a hidden feature approach could be developed and made plausible.

Charles Siewert's *The Significance of Consciousness* is a model of philosophical care: it proceeds slowly, venturing to address the most preliminary questions in the study of consciousness, refraining from a rush to grand theories. Its humble, conscientious approach is, I venture to say, Aristotelian in character. It's precisely what we need at this point in the investigation of consciousness.<1>

My comments will focus on the material in chapters 3-5. In these chapters, Siewert develops a thought experiment designed to make conscious experience "conspicuous by

its absence" (73). The thought experiment requires us to imagine a character (named "Belinda") who has a certain range of abilities but lacks a certain visual experience; Siewert's key claim about the experiment is that Belinda is, as described, genuinely possible. This claim is then used to criticize a number of theories each of which implies that conscious experience is identical with the possession of certain "manifest" features -- that is, features that we can empirically detect without extraordinary or invasive procedures. I concur with Siewert's judgement against those theories. He also mentions, however, the possibility of a "hidden feature" theory of consciousness. Such a theory is, he acknowledges, left unscathed by the possibility of Belinda. His discussion of it (4.9) is brief, however; he is content to raise a couple of intimidating difficulties facing any such theory and move on.

My goal here is twofold. First, I want to deepen Siewert's critique of the manifest feature theories he attacks by pursuing a number of ways in which one might attack the claim that Belinda is genuinely possible. Siewert himself spends some time defending that claim against attacks, but I want to consider some further attacks against it, including one which I have myself been tempted to endorse in the past, and show how they fail as well. Second, I want to explore the prospects for a hidden feature theory of consciousness -- which is the sort of theory I am presently inclined to endorse -- in light of the possibility of Belinda, the specific difficulties raised by Siewert, and, in general, Siewert's observations about consciousness and attempted theories thereof. In brief, I want to get clear on how his contribution to the literature constrains the development of such a theory. If my remarks are on the right track, his work has enabled me to get a better view of just how such a theory might work.

Here, then, is the game plan. In the first section, I lay out the key thought experiment involving Belinda and get clear about its relation to a hidden feature theory. In sections 2 and 3 I consider some ways in which one might attempt to show that the apparent possibility of Belinda is an illusion. Section 2 is devoted to what I call the "standard strategy" of debunking apparent possibilities; section 3 introduces a strategy that turns on the peculiarities of the "first person" approach that Siewert emphasizes. None of these work, so I concede that Belinda is possible.

Hidden feature theories can accept that Belinda is possible, but they face two serious difficulties. First, they rule out other apparent possibilities as illusory. Once one accepts the genuine possibility of Belinda, it becomes hard to motivate ruling out these other possibilities. Second, there remains the central question: how exactly are we supposed to determine a posteriori which hidden feature is *the* feature that makes for conscious experience?

The remainder of my discussion is devoted to thinking about what a hidden feature theorist can say about these two problems. In section 4 I argue that a hidden feature theory is in a good position to dismiss certain apparent possibilities as merely illusory. The final section turns to the more delicate issue of what we might know a priori about conscious experience that would make sense of an a posteriori identification with some hidden feature while being consistent with the genuine possibility of Belinda.

1. The Key Thought Experiment and Its Critical Impact

1.1 The Key Thought Experiment

The key thought experiment focuses primarily on two imagined characters, Belinda and Connie, who differ in the following key respects. Connie is an ordinarily sighted person on her right field of vision, but she has very poor vision on her left; she has visual experience of objects on that side, but matters are blurry enough to render her unable to report what is on that side beyond very rough estimates of size, shape, and so on. Belinda is like Connie with regards to her right field of vision, but she is entirely lacking in visual experience of her left field of vision. Nonetheless, Belinda exhibits a kind of blindsight; she is able to report reliably on objects in her left field of vision, reports matching Connie's in their degree of detail and accuracy. Further, Belinda can make these reports spontaneously, and she even has spontaneous, reflective thoughts about the character of these thoughts - about, for instance, their causal origin and role in her overall cognitive economy.

The case of Belinda is central to Siewert's discussion, as he uses her as a test for "consciousness neglect." More precisely, he stipulates that if a theory, allegedly about consciousness, implies that Belinda is strictly impossible, it is not, after all, a theory of consciousness at all -- not in the sense of "consciousness" that Siewert wants to focus on. This may seem like a terminological maneuver guaranteed to secure the possibility of the case, but it is ultimately innocent, as Siewert confronts at length the question whether consciousness in this sense should be neglected.

1.2 The Ease of Performing the Thought Experiment

The question is made easier to address because of the ingenious construction of the case. While Belinda is suited for criticizing a wide range of theories of consciousness, the case is evidently designed to require as little departure from actuality as is feasible. There are two ways in which the case is easier to imagine than many typical thought experiments in this area. First, it should be stressed that we are not asked to imagine "super-duper blindsighters" (p. 79) who have no visual experience but have abilities exactly similar to those who have the full range of conscious experience. Instead, we need only imagine Belinda, whose blindsighter abilities are relatively modest; for the contrast case we imagine Connie, who has the contrasting experience but whose abilities are, as it were, cut down to Belinda's size. Second, the case is made easier to assess by the fact that Belinda's lack of conscious experience has some typical behavioral effects. Belinda and Connie are not, after all, entirely alike with respect to their abilities. They differ in at least this respect: Connie can report truly that she has certain experiences whereas Belinda can report truly that she does not. We are not asked to imagine cases in which the difference in experience makes no difference to the subjects' behavior. There is no hint of

epiphenomenalism in this case. Our intuitions regarding the case are not likely to be confounded by trying to cope with the oddness of supposing that differences in experience are incapable of making a difference to behavior.

1.3 The Critical Impact of the Thought Experiment

To appreciate the significance of the thought experiment, it will be useful to introduce a modest classification of theories of consciousness. Let's distinguish reductive from nonreductive theories; manifest feature from hidden feature theories; and functional from nonfunctional theories.

Let's say a theory of consciousness is "reductive" just in case it implies that conscious experience is nothing over and above something which can be understood antecedently to understanding consciousness. A manifest feature theory is one that implies that conscious experience is nothing more than some feature that is "manifest," that is (in Siewert's terms), a feature "we can have warrant for thinking we have, without having to observe anything hidden literally inside us (for example, in our skulls)" (p. 107). Hidden features are those not manifest. Finally, we should distinguish those theories that imply that consciousness is nothing more than a collection of abilities, dispositions, or propensities from those that don't; call the former functional theories.

Those theories undermined by the possibility of Belinda are reductive accounts which imply that conscious experience is nothing more than the possession of certain manifest abilities. That is, they are reductive, manifest functional feature theories. Nearly every ability proposed by these theories as constitutive of consciousness is one that is shared by Belinda and Connie. As a result, those theories imply that the two cannot differ with regards to their conscious experience. Since they do differ in that way, however, we must either give up the theory that implies they cannot or reject the claim that such a pair of cases is possible.

Some of the theories Siewert criticizes do not admit of such straightforward criticism, since the abilities proposed to be constitutive are not shared by Belinda and Connie. For instance, one such theory (pp. 113-115) implies that conscious experience consists in certain abilities to acquire concepts: perhaps Belinda cannot acquire concepts based on her left-field blindsight that someone with left- field visual experience could. Siewert's attack on these theories takes the following general form: they find a genuine difference between Belinda and Connie, but the difference in question is one that is explained by the fact that they differ with regards to conscious experience; it is not one that can be used to explain their differing with regards to conscious experience.

I find Siewert's criticisms of these theories quite persuasive, and I do not want to defend them. What I want to do, rather, is defend a different sort of reductive theory of consciousness, one that makes conscious experience a hidden feature.

1.4 Hidden Feature Theories

Let me first sketch the kind of theory I find attractive and then consider how it is related to Siewert's discussion of the thought experiment.

As is now familiar from work by Davies and Humberstone (1980), Chalmers (1996), and Jackson (1998), one might hold that a particular predicate, say, "is F," is such that a competent speaker can know a priori just how features of the context of utterance determine what property is expressed, where she may be unable to know about those features of the context without empirical information. As a result, if she is ignorant of the relevant features of the actual context of utterance, her ability to assess whether certain descriptions using "is F" are genuinely possible is unreliable. Once the contextual information is in hand, however, she should be able to evaluate their possibility in a relatively a priori fashion -- that is, relative to her already having that key contextual information.

One considerable advantage of a "hidden feature" theory is that it allows the possibility of both Belinda and Connie. Belinda and Connie might differ with regard to such a hidden property, since we've specified the cases only with reference to manifest properties. More precisely, the way these two individuals have been specified leaves open various parameters, one of which might be relevant to explaining the difference between them. If such a hidden feature view of consciousness is correct, then *some* versions of Belinda are possible -- those in which she lacks the hidden feature. So such a theory would not stand accused of neglecting consciousness; it admits the central intuition Siewert exploits.

None of this is news to Siewert; he admits all this explicitly (p. 147). Nonetheless, he exhibits no enthusiasm for the approach. Instead, he raises two serious difficulties for it and confesses to having no view of how one might develop the approach.<2>

1.5 Two Problems For Hidden Feature Theories

The two difficulties he directs specifically to hidden feature theories concern, first, the apparent conceivability of other cases which would conflict with a hidden feature view, and, second, the question of how one could discover what hidden feature is to be identified as constitutive of consciousness.

As is well known, David Chalmers (1996) has argued that "zombies" are genuinely possible -- that is, it is possible for there to be creatures physically indiscernible from us but entirely devoid of consciousness. While he does not invoke zombies in his main argument, Siewert points out that a hidden feature view -- which presumably would make the hidden feature some sort of physical feature -- would face the challenge of the apparent possibility of zombies:

> Though the possibility of such "zombies" goes well beyond what I have asked you to contemplate in connection with blindsight, acceptance of the reflections I have invited may well leave you thinking that nothing ultimately rules out the possibility of a world such as Chalmers describes. (p. 147)

We need not even go so far as Chalmers-style zombies to make the needed point. We need only say this: consider any alleged hidden feature F and modify the description of Belinda's case so that she now has F. Is it still a possible case? That is, is it possible for her to be just as before with regards to her abilities, have F, and still lack conscious experience on her left side? It certainly seems possible, for any physical F you pick. So the hidden feature theorist faces a challenge: explaining away the appearance of *that* possibility.

The second difficulty is obvious enough:

> How do we decide *which* hidden features are, as a matter of *metaphysical* necessity, phenomenal features? Evidence that various forms of conscious experience will not occur in us, if certain hidden kinds of brain activity do not take place, and that such experience will occur in us if those kinds of activity do, presumably would not sufficiently justify the conclusion that *in no possible world* might some being lack brain activity of just that sort but talk noiselessly to itself, feel an itch, or see orange. ... We may say that there is a right answer to the question of how to state the hidden conditions of consciousness in a way that ranges over all possible worlds. ... But if God only knows what that answer is, and we have no idea of how to justify a claim to have found it ourselves, this does not help us much to understand the nature of consciousness. (pp. 147-8)

Plainly, one could consistently hold the view that some hidden feature is to be identified with conscious experience even while denying that we can ever discover which feature it is. McGinn's "mysterianism" about consciousness seems to be an example of this view (see McGinn 1991; the label is due to Flanagan, 1992). But such a view is likely to seem, at best, unsatisfying, and at worst a mere article of faith. In any case, I expect it is obvious from my earlier discussion of a priori knowledge of the relevance of contextual information to determining identity claims that this is not a view to which I am tempted. The view I want to defend is one which holds that we do know, a priori, just how contextual information could be relevant to determining the relevant hidden feature. I don't think it's easy to see how this might work, by any means. Indeed, the Belinda example imposes some important constraints on what the semantic story regarding the determination of the reference of experiential terms might be. In section 5 I will try to say something about how a hidden feature theorist could proceed here.

For now, however, I want to return to the possibility of Belinda, to strengthen the case for her genuine possibility.

2. Explaining Away Apparent Possibilities: The Standard Strategy

2.1 The Standard Strategy

There is a well-known strategy for explaining away apparent possibilities -- I'll call it the "standard strategy" -- and Siewert spends some time (section 5.3) criticizing its application to the case of Belinda. His target is not, in that section of the book, hidden feature theories; it is rather those theories that imply that *Belinda* is not possible. Here is how Siewert initially characterizes the strategy:

> In the water example, we are offered an account of why what it is said is not possible (i.e., that some water is not H2O) nonetheless *seems* possible, which is supposed to help support the claim that this is not possible. The account is that it seems possible to us because we confuse (or fail to distinguish) the superficial appearances whereby we judge that something is water with the property of being water, and so mistakenly think of a situation in which some non-H2O has that appearance, for one in which some non-H2O is water. (p. 158)

Thus put, however, it seems immediately obvious that the standard strategy cannot be applied to cases in which consciousness seems to be absent:

> But a similar account [to the one given for water] does not seem available in the consciousness case. For one cannot say that we are confusing the "superficial appearance" of conscious experience with the property of consciousness (its "real essence"), as one would perhaps say we might confuse the superficial appearance of water with the property of being water. Experience has no appearance, "superficial" or otherwise, in the relevant sense. (pp. 158-9)

"Appearance" is as historically burdened a philosophical word as any, but this much is certainly a steadfast association: if we think of "appearances" we are likely to think of sensory appearances, the way something looks, feels, sounds, etc. And in this sense it is quite clear that experience has no such appearance.

2.2 Consciousness and "Appearances"

This is hardly the end of the story, however. The standard strategy can be understood in a way that leaves out entirely talk of "appearances." The core of the strategy can be characterized schematically in the following way. Upon performing a thought

experiment, we find that a certain scenario seems possible. We are then tempted to assent to an *apparent possibility* (AP):

AP Possibly: Some individual has P but not C.

The strategy hopes to explain why we are tempted to AP in such a way that removes the temptation. It aims to do this by specifying some condition C* and making two claims about it. First, while AP is false, it is *genuinely possible* (GP) for some individual to have P but not C*:

GP Possibly: Some individual has P but not C*.

Second, our temptation to assent to AP is due to our recognizing the truth of GP and presuming, falsely, that C* is essential to C. That is, we make the *false presumption* (FP):

FP Necessarily, if some individual has C, she has C*.

The strategy, then, holds that we find AP plausible only because we have implicitly reasoned from GP and FP to AP. But since FP is false, that reasoning is defeated.

Now, the characterization I just gave is in important respects broader than the characterization given in the passage quoted above. Condition C* need not be in any sense an *appearance* of C. In some cases, of course, it might be an appearance that is falsely presumed to be essential; for instance, in the well-known case of identifying heat with molecular motion, condition C* happens to be something we can reasonably describe as the appearance of heat, namely, being such that it produces the sensation of heat. (The famous water/XYZ case is a bit different. There, the modal error is that of supposing, not that it is possible for something to be H2O without being water, but that it is possible for something to be water without being H2O. The false presumption is not that the appearance is *necessary* to water, but that it is *sufficient* for being water.) But it is plain from the way in which the strategy works to explain away the appearance of possibility that the *only* requirement on C* is that it be some feature not genuinely essential to C yet liable to be implicitly thought to be essential to it.

The point is not unknown to Siewert. He recognizes it (without emphasizing it) at a later point in the text (pp. 159-160). I stress the point, however, because its neglect can lead one to a premature despair over the prospects of explaining away apparent possibilities regarding consciousness.

2.3 A First Failed Attempt

That point aside, is the strategy one we can plausibly apply to the case of Belinda? Let "E" be the name of the visual experience that Belinda lacks and "A" the name of all those abilities common to both Belinda and Connie. Our question, then, is whether we can

explain why it seems true that there could be some individual with A but without E. Let's call this apparent possibility the *Apparent Possibility of Belinda* (APB):

> APB Possibly: some individual has A without having E.

One option, explored by Siewert (p. 160), is to identify the "appearance" of consciousness with some thought one might have about one's experience. Suppose we say, then, that APB is false, but GP1 is true:

> GP1 Possibly: some individual has A (and has E) without having the thought that she has E.

By itself, of course, this is not enough to debunk APB. We also have to make plausible the claim that the reason we find APB plausible is because we recognize GP1 and reason from it and the false FP1 to APB:

> FP1 Necessarily, if someone has E, she has the thought that she herself has E.

The problem, as Siewert points out, is that it just seems false that our temptation to APB is based on implicitly assuming FP1. If we explicitly disavow FP1 -- as Siewert does -- the temptation *remains.* One might insist that really, what explains our temptation to APB is our acceptance of FP1, that we are fooling ourselves somehow; perhaps our temptation to APB is a case of belief persistence. Maybe so, but that rather skeptical story about the causes of our temptation to APB is simply unsupported at this point. We should be as suspicious of it as we should be of any unsupported accusation of self-deception.

2.4 A Second Failed Attempt

Of course, there might be other ways of implementing the standard strategy. Might there be some other condition to play the role of C* in GP and FP?

I want to suggest a second option. I don't think it succeeds, but it is worth pointing out, as it has considerably more promise than the one just explored. Instead of the thought that one has E, let us consider this property: not having the thought that one doesn't have E. Or, let us go even further and consider this property:

> Not being such that one is disposed, upon careful introspection, to believe that one lacks E.

Call this ugly-looking property "B". Now consider GP2 and FP2:

> GP2 Possibly: some individual has A (and has E) without having B.

FP2 Necessarily, if someone has E, she has B.

On this second option, the suggestion is that GP2 is true: an individual could have A, and E, without having B -- that is, while being disposed, upon careful introspection, to believe that she lacks E. But we suppose, mistakenly, that FP2 is true -- that is, that anyone who has E would not be disposed, upon careful introspection, to believe that she lacked E. Maybe she would not be disposed to believe she had E, but, at a minimum, she would not be disposed to believe she *lacked* E.

This attempt to explain away our temptation to APB has the great advantage of imputing to us a presumption about essential properties (that is, FP2) which is something we likely do have a hard time disbelieving. It's well-known that philosophers have through the ages been tempted to some sort of infallibility thesis about self- knowledge. So it is quite clear that, for whatever reason, we are constantly tempted to believe that if someone introspects and has such-and-such beliefs about her mental states, certain *radical* errors, at least, are not possible. We can readily accept errors of milder sorts: we have no trouble with the claim that we can mistake a hot sensation for a painful one, for instance. We are also ready to accept errors due to hasty judgements, inattention, and distorting expectations. But here, we are concerned with a belief about one's experience formed upon careful introspection -- not based on hasty inference or authority or the like -- and the belief in question is not just a misclassification of the experience but a denial of the experience altogether. It is very hard for us to believe that someone might have E but believe, upon careful introspection, that she lacks E.

So we are, indeed, tempted to FP2. But this advantage of the strategy is at the same time its weakness: for it to succeed, it has to make FP2 something to which we are tempted -- so that it can explain our temptation to APB -- but also something which we can agree is false. If we simply agree to FP2, the strategy fails. Is there any way to make it plausible that FP2 is false despite the difficulty we have believing as much?

Perhaps. One might argue as follows. It's hard for us to believe that FP2 is false. But whatever explains our persistence in believing it, there are powerful reasons to disbelieve it. If FP2 is true, then there is presumably some logical connection between E and B, but it is hard to see how there could be. The experience E is one thing; not having a disposition to believe that one lacks it is quite another. There seems no incoherence in supposing that one has E while having a pathological disposition to believe, upon careful introspection, that one lacks E. Of course, the cognitive machinery of such a person is presumably lacking in a very serious way; one might insist on building into the notion of careful introspection the successful functioning of certain aspects of one's cognitive machinery. In so doing, one could produce a version of FP2 that is true. But when we find it hard to believe FP2 false, it does not seem to be the tautology we find so compelling. That is, what we find compelling is not the trivial claim that if one has E, and successfully attempts to discern whether or not she has E, then she will not, as a result, end up believing she lacks it.

This attempt to use GP2 and FP2 to undermine our temptation to believe APB faces two serious difficulties. The first is that the strategy is in an important way incomplete. Until we have a good story about why we're so persistently tempted to believe FP2, the view remains unstable: we will continue to suspect that once we've gotten a clear view of our temptation to believe FP2, we will see an error in the argument for its falsity. Now, there is one small suggestion that could be made at this point. We have, at present, no good idea what causal mechanism links beliefs about our mental states to the mental states themselves. Our ignorance of those mechanisms may encourage us to accept FP2 without good reason. In brief, our ignorance of the actual mediating mechanisms may encourage the view that beliefs about our own mental states -- at least some of them -- are not at all mediated. And without such mediation it is hard to see how errors are possible. Perhaps this account of our temptation to believe APB can be developed in a satisfying way.

The second difficulty, however, seems to me fatal. For the standard strategy to work, it must not only establish that GP2 is true, FP2 is false, and that we are tempted to believe FP2. It must also make it plausible that our temptation to believe APB is *due solely to our acceptance of GP2 and FP2.* But that just doesn't seem plausible. When we set out to imagine Belinda, it is not as if we just imagine someone having certain abilities and certain beliefs and dispositions to believe. Included in any adequate description of what is imagined is the experiences she has and those she lacks.

Or so, anyway, I am inclined to say. But someone might be suspicious of the claim, wondering whether the "first person" approach involved in imagining Belinda might impair our ability to describe accurately just what is being imagined. After all, if what is indeed being imagined is a situation where Belinda has the experience but is inclined to believe she doesn't, then, if we were to ask *her* what is going on, she would tell us that she lacks the experience. Similarly, one might suspect, if what is truly being imagined is her having the experience but believing otherwise, and we are imagining being in her situation, then, if we were asked what was going on, we would say that we lack the experience -- but we would be wrong.

This suspicion is, I think, worth following up.

3. The First Person Method and an Alternate Strategy

3.1 The First Person Approach and the Thought Experiments

The first two chapters of Siewert's book are devoted to defending a "first person approach" (p. 4). More precisely, they defend the "distinctiveness thesis," that is, the claim that

> one has a type of warrant for some of one's beliefs or claims, assertible using a first-person singular pronoun to attribute some experience (or attitude) to *oneself,* that differs from the type of warrant had (ordinarily, at

> least) for any beliefs or claims, whose assertions could constitute the attribution of some experience or attitude only to someone other than the speaker. (p. 6)

I find this thesis overwhelmingly plausible. My question concerns the role it plays in his overall discussion. It is obvious enough what role it plays in his claims about our *actual* conscious experience: many of those claims are said to have, in fact, this distinctive first person warrant. What is not obvious is how it could play any role in claims about merely hypothetical cases. What role, if any, does it play in the performance and/or evaluation of the thought experiments?

It's not clear to me that it has any role in the evaluation of those experiments, but Siewert does make clear that it has a role in performing them:

> [W]here I describe merely hypothetical situations in which someone has or lacks certain kinds of conscious experience, I would like you to conceive of being a person who has or lacks the relevant sort of experience. (p. 5)

What is distinctively "first person" about Siewert's thought experiments, then, is to be understood in terms of the instructions we are to follow in performing the thought experiment. When I imagine Belinda's case, I am supposed to imagine being just like her in experiential matters; I am to imagine having and lacking the same range of experiences while imagining, further, that I have the same range of discriminative abilities and so on.

Suppose this is in fact how we perform the thought experiment. The standard strategy for explaining away the appearance of possibility focuses on mistakes we, as philosophers performing the experiment, might make; but the first person approach suggests a different strategy -- one that focuses on mistakes that might be made by Belinda, or by us qua Belinda.

3.2 An Alternate Strategy

The alternate strategy being proposed needs careful handling. The first point to be clear about is that it offers a diagnosis of our temptation to believe APB, namely, that it's a result of two things: our projecting ourselves into her shoes and adopting her belief that she has A but lacks E, and her belief's being mistaken. So the diagnosis depends on two claims, which I'll call the "Mistaken Belief" thesis and the "Source of Temptation" thesis:

> *Mistaken Belief.* What we successfully imagine when we try to imagine Belinda is someone with a mistaken belief that she actually has A and lacks E.

> *Source of Temptation.* Our temptation to believe APB is due to our imagining being that person and adopting her belief about the (for her) actual situation.

If we indeed perform the thought experiment by imagining ourselves being Belinda, the Source of Temptation thesis may be right. If we are, as it were, basing our belief that APB is true on Belinda's authority, then, if her authority is undermined, our belief that APB is true is itself undermined.

Now, the first point we should be clear about is that the mere denial of infallibility is of limited dialectical force. One might argue as follows:

> Belinda believes that she lacks E. But this belief is fallible, since all beliefs are fallible. So it is at least possible that her belief is mistaken; hence, it is at least possible that our temptation to APB is based on a mistake. If we have powerful evidence of other sorts to believe a theory of consciousness that denies APB, we should suppose APB false, even in the absence of a positive story to tell about our temptation to APB.

This is all fair enough. But it doesn't go very far. What is really wanted here is an explanation of why we are tempted to believe APB, not just a justification of denying it. How could the alternate strategy sketched above provide us with that sort of explanation?

3.3 Should We Suppose that Belinda is Mistaken?

In a lengthy footnote, Siewert discusses some comments by Michael Tye that might be understood as an attempt to implement this strategy:

> [Tye] suggests that if you can imagine, and conceive of, some subject having the kind of visual representation he (Tye) identifies with the phenomenal character of visual experience, but without having that experience, you are mistaken. What you are *really* imagining is not what you say you imagine (that is metaphysically impossible), but rather someone who is oblivious to her visual experience in the way a distracted driver may be. ... I suppose Tye would say that I have misdescribed what I have imagined: what I have really done is imagine that Belinda is oblivious to how things look on her left -- she lacks a certain kind of thought about her left-field vision that she has about her right. (p. 353, n.1)

The suggestion made here on Tye's behalf perhaps counts as a version of the strategy I am presently considering. If I imagine being Belinda, and imagine being oblivious to how things look on my left, I may infer, mistakenly, that there is no way things look on my left, and hence that a case is possible in which there is no way things look to her left.

This strategy could be successful, it seems, if our actual imaginative act is well described as one in which we imagine being someone who is in a cognitive state that plausibly could mislead her as to her own experiential states. If Belinda is oblivious to her left-hand visual experience, she might be disposed to believe she lacks that experience, and that belief of hers might (in the manner of the Source of Temptation thesis above) affect our own theorizing.

But the strategy thus envisioned is unsuccessful, as it is not an adequate description of the actual imaginative act to say that we imagine Belinda being in an oblivious state. Siewert's rejection of the strategy is, I think, right on target:

> I am indeed conceiving of Belinda as someone who does not think she has (in my sense) phenomenally conscious visual experience of what is on her left. ... But that is not *all* I am supposing about her. I am also supposing that she does not have the phenomenal experience she does not think she has: not only does she not think it looks any way at all to her on her left -- it does not. And I am certainly not imagining her to be in a distracted-driverlike state; for I am not in the least tempted to confuse what it is like to be in that state with what it is like to be someone to whom things do not look any way at all. (p. 354, n.1)

The key point we should take away from the failure of this strategy is that it is not enough to point out that we might imagine ourselves being a person who either fails to believe she has E or believes falsely that she lacks E; we must also make it plausible that *our actual imaginative act* is describable as one of those.

Is the strategy then hopeless? While I think assimilating Belinda to a case of a distracted driver is quite a mistake, it seems to me there is a more radical suggestion worth considering. Unlike the suggestion considered and rejected above, where the source of error in Belinda's belief was already understood, subsumed under a familiar and specific case of error, the suggestion I want to introduce is that the manner in which we *generally* form beliefs about our own experiential states is the sort that can lead us to make errors of an unfamiliar sort. Let me explain.

First-person experiential beliefs -- by which I mean beliefs about one's own experiential states -- are typically noninferential. I do not infer that I have such-and-such experience; I simply form the belief spontaneously. Of course, something in fact causes me to have the belief; hopefully, the experience itself figures in the causal explanation. But my practice of forming beliefs about my own experiences in no way relies on any knowledge of that causal process. What, then, happens when I try to imagine being someone with certain beliefs about his own experiences?

In the actual case, we know that our first-person beliefs are spontaneously formed, not constrained by being inferred from other beliefs. As a result, one might argue, even if we should, when imagining being Belinda, infer that we have E (indeed, because it's implied by having A), we won't, because such beliefs aren't formed as a result of inference in the

actual case. To put the picture vividly, here is what happens, according to this line of thought, when I imagine being Belinda: I imagine having all the abilities A; I ought to infer that I have E, because having those abilities implies having E; but my ordinary practice of forming first-person beliefs will not allow such a belief; instead, I wait, as it were, to see if such a belief spontaneously appears in me. Since it need not -- because I am not actually, causally hooked up to Belinda's situation the way such spontaneous beliefs would require -- I never form the belief, and consequently infer I do not have E.

Of course, we *sometimes* infer first-person beliefs from other beliefs, so the position as stated above needs some qualification -- perhaps this: If I am tempted to infer a belief about my present or near-future experiential states, I will not actually succumb to the temptation (for any length of time) unless it is confirmed by arising spontaneously. In the imagined case, perhaps this is what happens: I might be tempted to infer that I have E, but because the belief does not also arise spontaneously, I do not give in to the temptation.

3.4 Difficulties with the Strategy

Although there is something attractive about this strategy, it faces three considerable difficulties, the second and third of which I regard as individually fatal.

First, and most obviously, this explanation is very speculative indeed: it requires that we presume a particular story about what, exactly, is going on when we imagine being Belinda, and it is not at all obvious that this story is correct. One may readily object that it is too crude to be realistic. When I imagine being Belinda, I also have other spontaneous first-person beliefs about my actual self; but I don't let those interfere with my imagined situation. (I spontaneously believe that I am male, but I don't suddenly find myself inclined to think Belinda believes this.) This shows, at the very least, that the story needs clarification and complication.

Even if the story is correct, or some version of it is correct, it is not at all clear that it helps. Suppose that our first person experiential beliefs are indeed such that, necessarily, we do not have them (for long) unless they arise spontaneously. The denier of Belinda's possibility is, nonetheless, going to be an advocate of some theory according to which the content of the spontaneous belief that I have E is the same as the content of the belief that I have A. If this is so, then why is it not transparent to us? What story are we going to tell about our failure to see this implication? If, when performing the thought experiment, I resolutely refuse to infer that I have E, why do I refuse? Someone might say: I am built that way; what is special about first-person beliefs is precisely their being ones that we cannot form without their being (typically anyway) formed spontaneously. Maybe so. But notice: this doesn't ultimately help, because there is another belief that I could infer, and should infer, from the fact that I have A, if the above story is correct. The other belief is this: *that if I should happen to spontaneously believe I have E, that belief would be true.* Even if I cannot, in performing the thought experiment, form the belief I have E, I surely

could form the counterfactual belief that if I were to form the distinctive first-person belief, the belief would be correct. Evidently, however, I feel no compulsion to infer such a thing from the fact that I have A.

A final reason for my not wanting to pursue this approach further is that I am doubtful that there is anything to be gained from a distinctively first person approach to these thought experiments.

3.5 The Irrelevance of the First Person Approach

Let me be clear: I find Siewert's "distinctiveness thesis" about the existence of a special kind of first person warrant utterly plausible. What I doubt is that focus on the first person makes any difference to the performance of the relevant thought experiments. As a result, my main complaint about the above strategy is with the Source of Temptation thesis -- the claim that our temptation to believe APB is due to our imagining being some person and adopting her beliefs about the (for her) actual situation.

There are three points I want to make here. First, it seems we could perform the thought experiment perfectly well without thinking of it as in any way especially "first person" in character. Instead of asking "Can I imagine being a person like Belinda, with her experiences, lack of experience, abilities, and so on?" I can ask simply: "Can I conceive of there being a person like Belinda...?" I am not at all sure what is added when asked to consider the situation from the first person. In any case, performing the thought experiment in the second, impersonal fashion yields the same result: APB seems true.

One might worry that the "impersonal" version could succumb to the following objection: what I end up imagining is being someone *other* than Belinda and ascribing to her various abilities, experiences, and so on. And in that case, we might confuse what we would be warranted, from the third person perspective, in supposing about Belinda, with what could in fact be the case about her. Siewert does tell us, after all, that he is concerned to avoid such confusions:

> Why do I make a point of asking you to focus on the first- person case? Because I think that if instead you consider only the third-person case -- that is, if you think only of instances, whether actual or not, in which someone other than yourself might be said to have or lack a given sort of experience -- you are liable to attend not to the difference between consciousness and its lack, but to the differences in behavior that would warrant your either affirming or denying that another person had conscious experience of a certain kind. (p. 5)

I agree that such confusions must be avoided, but we need not avoid them by any special method of imagining being Belinda, as opposed to imagining being an observer of Belinda. What we can do, instead, is not to imagine being anybody in particular: we can

simply consider the described situation and inspect it for signs of incoherence. That is how thought experiments normally work, after all.

The second point I want to make is this. To the degree that I *can* make any sense of performing the thought experiment in a distinctively first person way, the experiment actually seems to me to be problematic. Here's why. Suppose you are asked to imagine being someone in situation S, where S includes explicit descriptions of both experiential states and the rest of the person's situation. I take it that if I am asked to do this, what I do is simply: imagine having a certain set of experiences -- where the experiences I imagine having are not limited to those explicitly described in S but include as well those which may be supposed likely had by someone in that situation. ("Imagine winning the lottery just after you land the stressful but high paying job you've worked for most of your life. How would you feel?") The point I want to stress is that such imaginative acts are guided by one target: imagining having certain experiences. To imagine being a person is to imagine having such-and-such experiences; it is no more than imagining *what it would be like* to be such a person.

If the target of the imaginative exercise is limited to the imagining of having certain experiences, it hardly seems sensible to include as part of the requirements on imagining such a thing that you succeed in imagining that various extramental conditions obtain. Here is one way to make the point dramatic. First, imagine yourself being Descartes, sitting in his study, pondering his state of knowledge. Second, imagine yourself being Descartes, asleep in bed, dreaming that he's in his study, pondering his state of knowledge. What did you imagine differently? More precisely: If we ask "what was it like in the first case as compared to what it was like in the second case?" the answer is: they were exactly the same. The different extramental characteristics directed you to the same experiential states.

If I am right, however, then the first person method is unavailable as a means of evaluating APB. For consider: the imaginative act that would support APB would be an act of imagining not only being like Belinda in her range of experiences, but also in having her discriminative abilities -- the nature of which requires the holding of various extramental conditions. To drive the point home, we can ask: what is the difference between (i) imagining being Belinda, lacking E, but being in fact reliable with regards to her spontaneous beliefs about objects in her left hand field of vision, and (ii) imagining being Belinda, lacking E, and also being entirely unreliable with regards to her spontaneous beliefs about objects in her left hand field of vision? If there is no difference to what is imagined, then we obviously cannot use the success of the imaginative act as support for APB: we would have no good reason to think our success reflected on the possibility of someone lacking E while having A as opposed to the possibility of someone lacking E while not having A.

Someone might, in fact, try to use this point to attack APB, by saying that what Siewert and the rest of us have successfully imagined is not support for APB at all, since our imagining of Belinda is just imagining being Belinda, and we have just seen why success in such imaginings is not support for APB. I am not, however, inclined to take this line,

as I think it is clear that we find Belinda conceivable, and that in so doing we are not just imagining having her range of experiences.

The third point I want to make here is just that the distinctiveness thesis Siewert defends is quite compatible with what I am saying here. I do not see any clear way for the thesis that we often have a distinctively first person kind of warrant to motivate the claim that we need to perform thought experiments about consciousness in any special first person way. One might suggest that this special warrant somehow should attach to the beliefs we form to the effect that such- and-such situations (involving consciousness) are possible, but this seems just false to me. While there is a distinctive kind of warrant we have for (many of) our beliefs about our *actual* mental states, that kind of warrant does not seem to be warrant for any beliefs about what mental states are possible -- much less for any beliefs about what *combinations* of mental states and other states are possible. So I suggest we think of the key thought experiment in a very pedestrian fashion: we produce a description of a situation and consider a priori whether it is possible; our consideration is then constrained by nothing more and nothing less than what we know about the way our concepts work.

4. Hidden Feature Theories and Problematic Possibilities

4.1 Explaining Away the Apparent Possibility of Melinda

Let us take stock. So far, I've explored a number of different strategies one might use to explain away the apparent possibility of Belinda, to explain why we are tempted to believe APB. The verdict has been negative in each case. And indeed, the position I want to defend accepts APB. But it denies another possibility which may seem just as plausible. Suppose feature F is the hidden feature that makes for consciousness (or, more specifically, for that particular visual experience lacking in Belinda). In fact, let *Melinda* be a hypothetical character who is just like Belinda with regards to her discriminative and other abilities and her lack of conscious visual experience of her left hand visual field, but with one difference: Melinda, by stipulation, has F. The hidden feature theorist faces the task of explaining away the apparent possibility of Melinda.

Melinda has A, has F, and lacks E. Is she possible? She certainly seems to be -- for a wide range of substitutions for "F". So the hidden feature theorist has to explain away our temptation to believe (APM):

> APM Possibly: Some individual has A and F but lacks E.

How might we explain away our temptation to APM? We might want, here, to go back and re-examine the strategies previously explored in considering APB; perhaps they will work for APM even though they failed for APB. I don't see any hope in that direction,

however. What I want to do is introduce a distinct strategy and make clear how it relates to the standard strategy.

A hidden feature theory naturally suggests the use of the standard strategy, since the famous examples of natural kind cases lend themselves readily to such diagnosis. As a result, one might feel compelled to use that strategy in defending a hidden feature theory; but this would be a mistake. I want to emphasize that we can exploit a hidden feature theory without having to suppose that anything plays the role of a presumed essential feature of E.

The standard strategy requires that we diagnose the intuition of possibility as the result of a false presumption. It's clear enough, I think, that any way of explaining away an apparent possibility must rely on our implicitly making an error. But there are different ways of making errors -- some more crude than others. The kind of error the standard strategy imputes to us is the positive belief that a certain feature is essential to consciousness. But we might be able to explain away an apparent possibility by supposing only that we are *ignorant* of something -- for instance, we might simply be ignorant of the fact that F is identical with E. The point I want to stress is that this ignorance alone is enough to explain the appeal of APM.

4.2 The Inability to Perceive Incoherence

The diagnosis I want to offer can be articulated in two steps: (i) An intuition that something is possible can be generated simply by the fact that we are unable, when considering the situation, to detect any incoherence; (ii) If we don't know that feature F is identical with E, our inspection of the case of Melinda will not enable us to detect that it is, in fact, incoherent.

The strategy is quite simple. We might illustrate it with an obvious example. Someone who doesn't know that "Amy" and "Clara" name the same person will be very tempted to believe that it's possible for Amy to have been born earlier than Clara. Notice that in no sense do we need, here, to specify something that is presumed essential either to Amy or to Clara. The same kind of explanation is available, if a hidden feature theory of consciousness is correct.

One might fairly point out here that with the example of Amy/Clara, the person who is tempted to believe it is possible for Amy to have been born earlier is making the false presumption that "Amy" and "Clara" refer to distinct individuals. That's right; what I want to stress is, however, that this sort of false presumption is the sort one can make quite easily. One standardly presumes that "A" and "B" name distinct things if the names themselves are distinct. As a result, the diagnosis on offer doesn't take on the burden of positing any special moment in the psychology of those who find the case to be possible.

4.3 The Remaining Burden

Siewert recognizes (briefly) the sort of strategy I am proposing on behalf of a hidden feature theorist:

> I am not saying that there is no explanation of my not apprehending the alleged impossibility. After all, we might explain it somewhat along these lines. On reflection, the description I have given of Belinda seems intelligible to me and without inconsistency, and I cannot find any good reason to think it describes a strictly impossible situation, so I do not believe it is impossible, and maybe even regard this as giving me reason to believe it is possible. (p. 163)

As Siewert rightly points out, to implement this strategy successfully one needs to tell a further story that tells us why we should believe the situation is impossible and which makes it plausible that the reason we have (thus far, at least) failed to detect the impossibility is that we have not been aware of that reason:

> But *that* explanation, as it stands, does not help at all to make the case for impossibility. Of course, it might contribute to that case, if one could find a good reason why I should believe the situation impossible, and then one could explain my failure to appreciate the metaphysical impossibility of Belinda by explaining why I failed to appreciate this reason. But this just sends us back to the problem that one has made this strong claim of impossibility, in the face of appearances to the contrary, but has given us no good reason to accept it. (p. 163)

Insisting that a given case is impossible without giving some story about how we could discover its impossibility seems desperate. So long as we have not offered a story about how we might go about discovering, a posteriori, which feature is to be identified with E, holding onto a hidden feature theory may seem a mere article of faith.

This charge is not quite fair, as one might have evidence for the claim that E is to be identified with some hidden feature without having any idea how we might gain *further* evidence regarding *which* feature is E. Further, one might insist that the demand for such a story in advance about a posteriori discovery is unwarranted, as we cannot, in general, tell a priori how we could make such discoveries. This latter line has some initial plausibility, as one might assimilate discoveries of identities to all other empirical discoveries, holding that they turn on an unruly collection of empirical and explanatory factors, including establishing correlations, preserving simplicity, and so on.<3>

Whether the charge is fair or not, it would certainly be more satisfying, at least, if we could give a story specifying what we know, a priori, about the relevance of empirically discoverable aspects of the actual context to the determination of the reference of "E". Now, I am hardly going to give a detailed story in the next section, but I do want to

consider some initial moves, some difficulties, and how I think, in broad outline, the story might go.

5. Hidden Features and A Posteriori Identification

5.1 What We Might Know A Priori about the Reference of Experiential Terms

Hidden feature theories are easily associated with the standard examples of natural kinds and the determination of the reference of natural kind terms. But we should not be too hasty in what we presume about the sorts of contextual information that we might exploit in giving a semantic story about the determination of the referents of terms for experience. The contextual information that we know a priori to fix the content of such terms need not be information about the "appearance" of experience. It need not even be information about conditions that we typically use as means of identifying such conscious states. This is a good thing, of course, because, as we've seen a few times above, we typically don't *infer* the presence of experience in the first place; hence, there are no conditions typically used as indicators or signs of it either.

The sort of theory I favor will include some story about the semantics of terms like "E". Now, I've emphasized that this story need not take the following form:

> E = that property P such that in the actual world P explains the superficial appearance of E.

We can abstract away from the talk of the appearance of E here and consider the following schema:

> E = that property P such that in the actual world P meets condition C.

Condition C would then need to be spelled out in some plausible way. Such a proposal should be understood as an articulation of what we may know a priori about the way in which the reference of "E" is to be determined. The primary challenge facing a hidden feature theory is specifying C in such a way as to give us a plausible conceptual truth about "E."

Before we move to consider what C might be, it will be useful to settle some terminology. If a predicate "E" works in the way being suggested, we should distinguish what we may call, following Chalmers (1996), the "secondary intension" of "E". This is the reference of "E" given its actual context of utterance. So, on the sort of hidden feature theory I am concerned to defend here, the secondary intension of E is just F. We also need some way of talking about what is in common between two uses of "E" in two different contexts of utterance, where there is more in common than just sameness of orthography. Intuitively, we may characterize this common element as the two uses'

being governed by the same normative rule articulated in a claim of the above form. Call this common element the "concept" of E. We may say, then, that a claim of the above form gives us a truth about the concept of E, and this same concept may be found in two beliefs, both expressed as the belief that one has E, where the two beliefs differ in secondary intension.

5.2 The Actual Explainer of the Associated Abilities

So, what might condition C be? It can't be the "appearance" of E, but it might be a condition that is otherwise *associated* with E. Siewert has already done us the favor of detailing an associated set of conditions -- namely, those possessed by both Belinda and Connie, the discriminative abilities, the ability to have spontaneous reflective visual thoughts, and so on, all of which I have grouped together as "A." So one option is to say that "E" refers to whatever actually explains the possession of those abilities:

> (1) E = that property P such that in the actual world the possession of P causally explains why those who have A have it.

A few initial points about this proposal are in order. First, there is, as with all suggestions of this type, a threat that there may be no unique property in the actual world that plays the relevant explanatory role. Perhaps in one (actual) person with A, F explains the possession of A, while in another (actual) person it is G that explains it. In that case, according to (1), there is no such thing as E: it fails to refer because of the failure of the uniqueness assumption.

There are various ways of reacting to this, none of which I want to endorse or defend here: one might accept the possibility; one might complicate (1) so as to set up a default reference in case of the failure of uniqueness, including perhaps a default clause specifying some relativization to species type or the like; or one could modify (1) so as to let E be either that unique property or that property definable by disjoining every property that plays this explanatory role in the actual world. Which of these might be a plausible move is a good question, but I think we would do better first to assess the overall plausibility of proposals like (1). So I am going to set aside these issues and presume henceforth that any uniqueness presumptions are satisfied.

We should note that this proposal relies on the assumption that there is, in the actual world, no one like Belinda. If Belinda is an actual case, then, there is presumably a feature F she has which explains her possession of A. Since we're waiving uniqueness issues, we must suppose she shares this feature with everyone else who has A; as a result, her feature F will be identical with E. Hence, if she is actual, she has E -- but of course she does not have E. Fortunately, this assumption that Belinda is not actual is not one that it seems unreasonable to make.

Now, is (1) plausible? It allows the possibility of Belinda, since it allows that there is a *possible* person like Belinda, who has A but lacks E, but also lacks the key feature F. What explains this possible person's abilities is some distinct hidden feature. So we can acknowledge Belinda. We can't acknowledge Melinda, of course, but we can explain why she seems possible: unless we have already figured out that F is identical with E, we will not be able to tell that her situation is impossible.

Unfortunately, (1) will not do. Suppose that Belinda has the concept of E, so that, at a minimum, she knows that "E" should refer to the actual explainer of the possession of A. Now suppose she considers the question as to whether she has E. Given (1), she should conclude that she does: she could note that she has abilities A and presume that she has some feature explaining those abilities; she should then conclude (waiving issues of uniqueness) that she has whatever feature actually (that is, indexed to her world) explains the possession of A is one she has. This conclusion she would express as the belief that she has E.

Of course, in so concluding, the content of the belief she forms would have a secondary intension different from that of the belief someone in this world might form. Belinda's use of "E" would refer to some property other than F. Still, her belief that she has E is recognizably similar to one that we might form in this world. Our assertion that we have E would share a common concept with Belinda's assertion that she has E.

Now, the reason (1) fails is that it implies that Belinda must, if she is rational, considers the question, and has the concept of E, conclude that she has "E", that the thought she has using the concept of E is true. But she plainly need not conclude this. She can have the concept of E without seeing the need to conclude that she has E. But if (1) is correct, she would have to see such a need. So (1) is false.

The lesson here is important. Whatever account we offer of the concept of E, we must take care to ensure that we avoid the result that Belinda would likewise conclude that she has E -- even if the secondary intension of her conclusion differs from that of beliefs similarly expressed in our, actual world. And this point raises a red flag: since we don't know a priori which hidden feature is identical with conscious experience, what we know a priori about how the referent of "E" should be determined seems to be a set of rules that Belinda could follow equally well, and conclude, just as we do, that she has E.

I want to propose another semantic story about what we know a priori about the determination of the reference of experiential terms; and I will argue that it escapes the above described problem.

5.3 The Actual Explainer of First Person Beliefs and the Associated Abilities

One might think that it is not surprising that (1) fails, since it is in no way sensitive to one aspect of consciousness that seems central to our understanding of it, namely, its being such that we standardly form first-person beliefs about it in a noninferential fashion.

Of course, I was at some pains to point out above that the contextual features that fix the reference of "E" need not be features that we normally use to find out whether something has E. So the fact that (1) fails to reflect our actual typical means of finding out if someone has E is no demonstration of its failure. Nonetheless, since (1) does fail, one might suspect that it would *help* to produce a semantic story that does reflect our actual typical means of finding out if someone has E.

How might such a story look? We might offer this:

> (2) E = that property P such that: in the actual world, for every person who noninferentially forms the first person belief that he has E, that person has P and P explains his formation of that belief.

As it stands, (2) seems to be a terribly impoverished account of what we might know a priori about the reference of "E". We can, however, combine (1) and (2) in the following way:

> (3) E = that property P such that: in the actual world, for every person who noninferentially forms the first person belief that he has E and has A, that person has P and P explains his formation of that belief and his possession of A.

The advantage of (3) is that it accommodates our sense that both the noninferential character of our typical beliefs about our own experiences and the possession of A are conceptually central to E. It is (3) that I want to explore as a promising proposal (albeit oversimple in its present form) for what we might know a priori that could enable us to discern which feature is to be identified with E.

Now, how promising is (3)? What is startling about (3) is its use, on the right hand side, of the notion of the belief that one has E. If we think of (3) as a definition or analysis of the concept of E, we are likely to complain that it is *circular* in the sense that it tries to analyze one mental concept in terms of other mental terms. But what, exactly, is supposed to be wrong with such circularity in this context? Bear in mind that (1), (2) and (3) are here presented as articulations of what is known a priori by anyone who possesses the concept of E -- more precisely, what is known a priori which enables one to determine, using contextual information, the reference of "E". There is no straightforward objection to such an articulation being "circular" in the sense described. So long as (3) is *true*, and so long as it is something that a speaker can know a priori, it can do its work in helping us identify the hidden feature F as identical with E. So circularity *per se* is not an objection. Still, there may be a related objection.

The related objection is that (3) doesn't tell us enough about what it is for something to be a belief that one has E in order for us to have any confidence that the account of E we get from (3) is plausible, or even coherent. Until we can say something about what distinguishes the belief that one has E from other beliefs, the account is incomplete. If we tried to make use of (3) on its own to characterize the belief that one has E, we end up with something unsatisfying: the belief that one has E is just the belief that one has that property which, in the actual world, explains the possession of A and the noninferential formation of the belief that one has E. Repeated substitutions of the right hand side of (3) into the place of "E" in "the belief that one has E" yield an infinite regress.

The objection here is quite important, and it appears in Siewert's book when discussing (4.5) a distinct (but plainly similar) proposal, namely that what makes for conscious visual experience is just *thinking* that one has it. To illustrate the absurdity of this view, he compares it to the view that the only thing that makes a shoe an "ultra-shoe" is that it can cause the wearer to believe that it is an ultra-shoe. An amusing dialogue makes plain just how silly the view is:

- How do you like my new ultra-shoes?
- I see your new shoes--but what makes them *ultra*-shoes?
- Well, being an ultra-shoe is a lot like being a shoe, in fact, it's *precisely* the same as being a shoe, save in this respect: ultra-shoes have the capacity to make the wearer believe they are ultra-shoes.
- What's so special about that? Any shoes can cause you to believe that you're wearing shoes.
- Of course, but ultra-shoes don't merely have *that* capacity. They cause you to believe not just that you're wearing shoes, but that you're wearing ultra-shoes. And plain old shoes can't do that.
- But there's still no difference, unless there's a difference between believing you're wearing shoes and believing you're wearing ultra-shoes. And what's the difference between those beliefs?
- I've already told you: to wear ultra-shoes is simply to wear shoes that make you believe they're ultra-shoes. And *that*'s what I believe about my ultra-shoes that regular shoes are powerless to make me believe about them: that they're ultra-shoes. (p. 132)

I concur with Siewert that the ultra-shoes theorist here has failed to draw a distinction between shoes and ultra-shoes. Is the advocate of (3) bound to fail in the same way?

5.4 The Concept of E

The first thing I want to point out is that (3) is not subject to exactly the same problem that the ultra-shoes theorist is stuck with. The advocate of (3) can, indeed, tell you something about the difference between having E and not having E. It's not *just* the difference between having a propensity to believe you have E and not having the

propensity. Since E gets identified a posteriori with some feature F, the advocate of (3) can add that the difference consists in having F or not having F. So there is an advantage over the unfortunate ultra-shoes theorist.

But there is, still, a puzzle one can raise: what is the difference between *believing* that one has E and not believing it? It can't be explained in terms of believing that one has F, since, on this view, the belief that one has E is something you can have while remaining ignorant of the hidden feature F. So it looks like this approach still needs to explain what it would be to have that belief. What can be said?

Given the semantic story (3), we can hardly distinguish the belief that one has E from other beliefs by reference solely to which property one represents oneself as having, since the advocate of (3) has to distinguish the belief that one has E from the belief that one has F even while identifying E and F. So the difference between those two beliefs presumably is a matter of differing "modes of presentation" or the like. Now, this phrase -- "mode of presentation" -- may suggest an *appearance* of E, by which means one identifies instances of E; and of course we've seen that no such appearance is available. But there is a way of making sense of "mode of presentation" which makes no commitment to any such appearances. We may say that the belief that one has E requires the use of the concept of E, where that concept is essentially tied to some cognitive or inferential role that differs from the belief that one has F. Presumably, if (3) is true, that role can be characterized as an implicit acknowledgment of (3). Someone who has the concept expressed by "E" is someone who is prepared to move from the information that a given property P is the actual explainer of beliefs using that concept to the conclusion that P is identical with E.

But that can't be all there is to having the concept expressed by "E". The problem is that if we accept this as a complete account of the concept of E we are forced to acknowledge that (a slight variation of) Belinda will necessarily conclude (upon considering the question) that she has E. Again, the secondary intension of her belief will differ from the secondary intension of the beliefs we might have about E, but the belief will share with ours the concept of E. And that is an unacceptable consequence.

To see this point clearly, let's back up and make more explicit the complete account of the concept of E suggested by my above remarks. Then I'll consider the (slightly modified) Belinda scenario and show how the account of the concept of E is inconsistent with the conceivability of that scenario.

Let "M" be a variable that ranges over concepts. The account suggested by my above remarks may be set out thus:<4>

> (4) The concept of E = that concept M such that someone S who possesses M accepts as primitively compelling inferences of this form:
>
> > My having property P is actually responsible for my possession of A and for my noninferential formation of the

belief that I have M.

Hence, my belief that I have M is made true by my having P.

In other words, if we take (3) as the whole story about the concept of E, we may individuate that concept solely by its playing that privileged inferential role.

I turn now to the slightly modified Belinda case. As before, Belinda has A and lacks E. Further, let's suppose she has some concept -- call it M -- which meets the above conditions. In other words, if (4) is true, M is the concept of E. Finally, let's suppose that, on one particular occasion, she introspects and as a result forms, in a noninferential fashion, the belief that she has M. This seems to me to be a possible situation: there is nothing incoherent about the concept thus delineated and nothing about her having A and lacking E that would prevent her from either having the concept or coming to form the belief that she has M. But did she thereby, necessarily, form the belief that she has E?

No. Not necessarily, anyway: I don't want to rule out the possibility of her genuinely forming the belief that she has E, in a noninferential fashion, and its being a false belief. That might happen. But what I do want to reject is the claim that if Belinda fits the above description, she must be using the concept of E, that the belief she forms must be fairly characterized as the belief that she has E. It is not enough to note that the belief she forms must be one with a different secondary intension; it is plainly not necessary that the belief be one using the same concepts as our actual beliefs about E. In other words, while Belinda's concept M may share some features with our concept of E, there is no level of sameness of semantic type of which they both are instances.

One way to see this is to consider what we would say if we were to discover that we were in Belinda's position. If we were to find ourselves spontaneously forming beliefs attributing some property to us, and then we discovered that P was the property responsible both for those beliefs and our possession of A, would we feel compelled to conclude, on that basis alone, that we had E? Plainly not. Suppose, during a particularly reflective moment while thinking about consciousness, you noninferentially form the belief that you have some property that explains your possession of A; on its heels follows the thought that this property could explain other aspects of your mental life as well, including, perhaps, the belief just occasioned. Finally, you spontaneously form the belief that you have some property that explains your possession of A as well as the very belief which you are presently having. Given the content of this last belief, you are prepared to infer that, if F is the property responsible for your having A and for your formation of this last belief, this last belief is made true by your having F. But this last belief is surely not fairly characterized as the belief that you have E. Indeed, if you were Belinda, and knew you lacked E, you could still go through this process and remain free of cognitive dissonance.

So (4) cannot be the whole account of the concept of E. Of course, (3) may be true nonetheless; it is important, however, that we recognize its incompleteness. What is missing?

The answer is simple: *actual sensitivity to E* (that is, F). Part of what makes a concept the concept of E is the fact that its possessor is able to use it in recognizing instances of E in himself. Adding this element to (4) we get this:

> (5) The concept of E = that concept C such that someone S who possesses C satisfies both of the following:
>
> (i) S accepts as primitively compelling inferences of this form:
>
> > My having property P is actually responsible for my possession of A and for my noninferential formation of the belief that I have C.
> > _____
> >
> > Hence, my belief that I have C is made true by my having P.
>
> (ii) S is disposed to recognize as belonging to a common type different instances of F in her own person, and such recognition will (under appropriate circumstances) give rise to her forming in a noninferential fashion the belief that she has C.

Given (5), it is easy to avoid the consequence we were trying to avoid, namely, that Belinda must, if possessed with the concept of E, be inclined to conclude she has it. She need not be inclined to conclude this, since she will not recognize instances of F in her own person, and hence will not form, noninferentially, the belief that she has E.

Of course, the explication of the concept of E given in (5) is not something we could give a priori, since we don't know a priori that F is the hidden feature to be identified with E. But I don't see any reason to insist such explication proceed exclusively in an a priori fashion. Let me be clear: I insist that we know a priori how we could move from contextual information to knowledge of the reference of "E". But knowing this much a priori is compatible with (5): we could know clause (i) of (5) a priori but fail to know clause (ii) a priori. More precisely, we could not produce (ii) as it is expressed above, since we don't know that "F" and "E" corefer, but we know what is expressed by (ii) because we in fact know how to recognize instances of E (viz., instances of F), and we know a priori that part of what individuates the concept of E is the concept-possessor's ability to recognize instances of E in his own person.

The proposal I'm offering here is reminiscent of (and partly inspired by) Loar's account of phenomenal concepts as a species of "recognitional" concepts:

> Phenomenal concepts belong to a wide class of concepts that I will call recognitional concepts. They have the form 'x is one of *that* kind'; they are type- demonstratives. These type-demonstratives are grounded in dispositions to classify, by way of perceptual discriminations, certain objects, events, situations. Suppose you go into the California desert and spot a succulent never seen before. You become adept at recognizing instances, and gain a recognitional command of their kind, without a name for it; you are disposed to identify positive and negative instances and thereby pick out a kind. (Loar, 1997, p. 600)

As Siewert notes, if the concept of E is a recognitional concept in Loar's sense, a subject's possession of the concept does not ensure that she can know a priori the ways in which the concept of E relates to other, nonrecognitional concepts:

> Having that recognitional concept does not, as we might say, give me the information I need to answer this question: in which "possible worlds" are various statements employing that concept true? So, I may employ this recognitional concept in thinking about a certain kind of visual experience without its seeming to me there is any essential connection between what it allows me to pick out and what is described in ways manifesting my possession of the other, discursive concepts: thinking with recognitional concepts does not put you in a position to understand and assess connections of that kind. (p. 162)

Obviously, the view that the concept of E is a recognitional concept in this sense makes it easier to accept any given claim identifying E with some physical or functional feature: any alleged possibility in conflict with the claim can be dismissed as seeming to be possible only because of the recognitional character of the concept of E.

We can get a better grip on this point, I think, by considering an experiment Loar envisages by way of illustrating the character of recognitional concepts:

> We can imagine an experiment in which the experimenter tries to determine which internal property is the focus of her subject's identifications: 'again',... 'there it is again'. There seems no commonsensical implausibility -- putting aside foundational worries about the inscrutability of reference -- in the idea that there is a best possible answer to the experimenter's question, in the scientific long run. (Loar, 1997, p.601)

Suppose that the only internal state of the subject which could be the cause of the subject's declarations is neurological state N. If we were to ask the subject to consider whether it was possible for her to have N without having "that state" she was talking about earlier, she would be unable to discern any incoherence in such a scenario and, hence, likely suppose the case to be possible. If a concept is purely recognitional in this fashion, then it is easy to explain away apparent possibilities involving it.

I might, then, be able to think about "things of that kind" without being able to discern, a priori, any conceptual links between things of that kind and what I conceive of in a nonrecognitional fashion. Siewert's complaint about this is not that such concepts are impossible or never occur, but that our actual concept of visual experience is not that "cognitively primitive":

> It is not the case that the concept I use in thinking of visual experience in connection with the [Belinda and Connie] thought-experiments gives me no competence to assess the presence or lack of necessary connections between what it picks out and what is described in other terms (assuming we have such competence at all). Suppose someone proposed that there is nothing essential to what I am thinking of when I speak of phenomenally conscious vision beyond what any phototropic organism possesses. If I balk at this, and reply that it is at least possible that a sunflower, say, has no conscious visual experience, why can't this person object that since I am thinking about the phenomenal character of visual experience employing my merely recognitional concepts, I have no business speaking out on the possibility that something with a simple phototropic response may lack visual experience? Perhaps he will insist that it is metaphysically impossible that things might not look any way at all to the sunflower. If this is absurd, it indicates that the concept I have of visual consciousness and I use in first-person thought does include some ability to assess relations of necessity and possibility. So one cannot reasonably explain my supposed failure to understand that someone like Belinda could not possibly fail to have visual experience of a patch of light, by saying that in thinking about visual experience, I am employing a merely recognitional concept unsuited for the business of thinking about what is possible and what is not. (p. 162)

If our concept of E were as primitive as the recognitional concept investigated in Loar's imagined experiment, then for just about any hidden feature you can think of, it could turn out that E is identical with that feature. If it turns out that the only internal state that could be causally responsible for one's belief that one has E is the state of one's cranium being infested with termites, should we conclude that having visual experience of that degree of acuity on one's left hand side is nothing over and above having termites in one's skull?

Compare the case of the subject in Loar's imagined experiment; if the experimenter were to discover that the only cause of her "there it is again" was having termites in her skull, it is not so obviously absurd to conclude that having termites in one's skull is indeed what she was picking out -- precisely because her "there it is again" is so devoid of conceptual wealth. But our concepts of experience don't seem to be like this at all. If we were to discover that termite infestation was the only cause of our beliefs ascribing E to ourselves, we would, I venture, feel compelled to conclude either that those beliefs were uniformly false (eliminativism) or insist that there must be some mistake in the data that led to the conclusion that only termite infestation could be the cause.

So I agree with Siewert that we cannot simply treat our concepts of experience as bare recognitional concepts. But we can take advantage of Loar's notion in a more limited fashion. If (5) is right, the concept of E is to some degree recognitional (because of the second clause), but there remain considerable a priori limits on what *could* be the reference of "E". I take this to be a serious advantage of the general approach exhibited in (5): we can leave a role for recognition even while making it intelligible both how we could discover a posteriori a feature to be identified with conscious experience and how we can know a priori some limits to those identifications. Without the latter, any promise of physicalistic respectability seems a matter of faith; without the former, we don't do the concept justice. With both we might just be able to fit conscious experience into a physicalistic world.

Notes

<1> All uncited page references are to Siewert (1998).
<2> Siewert raises a third difficulty as well, but it is not one that I am concerned with here, as it is only a difficulty for hidden feature theories which also insist that the hidden feature in question is a functional one. The view I'm exploring makes no such presumption.
<3> I should acknowledge here a second source of possible dissatisfaction with my way of dealing with APM. Even after someone has determined that F is identical with E, she might find APM plausible -- when understood as a claim about *epistemic* possibility. After all, it could have (as we say) turned out that the actual world is such that the reference of "E" was something else. In that sense, APM is true: precisely because we can't tell a priori that APM is false, it is possibly true in the epistemic sense of "possibly true."

Why should this epistemic possibility cause anyone unease? We might reason roughly thus: the epistemic possibility of someone having F without E shows us something about the space of genuinely possible worlds, even if there is no genuinely possible world fairly describable as one in which someone has F without having E. If so, then, this implication of the epistemic possibility may be taken to have further, important implications for the nature of consciousness. (This, in effect, is a key step in Chalmers' overall arguments for the failure of consciousness to supervene on the physical; see Chalmers, 1996, 134.) I don't have the space here to explore this adequately, but I will just say that I reject the initial move from epistemic possibility to an implication for the space of genuinely possible worlds. I develop this point somewhat in my "Conceptual Analysis, Circularity, and the Commitments of Physicalism" (forthcoming).

<4> My formulation here follows the style of Peacocke (1992).

References

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

Davies, M. and Humberstone, I. L. (1980). Two Notions of Necessity. *Philosophical Studies, 38,* 1-30.

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

Jackson, F. (1998). *From Metaphysics to Ethics.* Oxford: Oxford University Press.

Loar, B. (1997). Phenomenal States. In N. Block, O. Flanagan, and G. Guzeldere (Eds.), *The Nature of Consciousness.* (pp. 597-616). Cambridge, MA: MIT Press.

McGinn, C. (1991). *The Problem of Consciousness.* Oxford: Blackwell.

Peacocke, C. (1992). *A Study of Concepts.* Cambridge, MA: MIT Press.

Siewert, C. (1998). *The Significance of Consciousness.* Princeton: Princeton University Press.

Witmer, D. G. (Forthcoming). Conceptual Analysis, Circularity, and the Commitments of Physicalism. *Acta Analytica.*