**Self-Knowledge**

Aaron Z. Zimmerman

(forthcoming in *Philosophy Compass*)

**Abstract**: Recent philosophical discussions of self-knowledge have focused on basic cases: our knowledge of our own thoughts, beliefs, sensations, experiences, preferences and intentions. Empiricists argue that we acquire this sort of self-knowledge through inner perception; rationalists assign basic self-knowledge an even more secure source in reason and conceptual understanding. I try to split the difference. Although our knowledge of our own beliefs and thoughts is conceptually insured, our knowledge of our experiences is relevantly like our perceptual knowledge of the external world.

## I. An Overview

Self-knowledge has been a central issue in western philosophy ever since the discipline's inception. Indeed, according to Pausanias, "Know yourself!" greeted Greeks of all stripes as they entered the famous Temple of Apollo at Delphi. In his *Lives of Eminent* importance of self-knowledge reached its pinnacle with Plato when he argued that a great deal of mathematical and conceptual information can be gleaned from introspection. To gain insight into the basic categories into which things are naturally sorted we need only recollect facts buried in us at birth.

Even as it is ordinarily conceived, self-knowledge is often particularly difficult to acquire. Because we lack critical distance from ourselves, people regularly overestimate

their own intelligence, beauty and likeability.[1] We all know how hard it is to admit that we are angry for no good reason or that we feel disappointed by the behavior of a friend or family member.[2] And how can someone really know whether she is courageous or resilient until she is tested by adversity? The quest for many kinds of self-knowledge therefore looks to be a lifelong project that requires a great deal of experience and reflection. Of course, we typically know more about our own actions than do other people who have their own lives to lead, and when we develop our self-conceptions we are aided by introspective knowledge of our own aspirations, thoughts, and fears—information that other may lack. Still, as Gilbert Ryle (1949, pp. 169-79) emphasized, to acquire knowledge of my character, personality and abilities, I must perform the same kinds of inference that others use in coming to know me. I simply have more data on which to base these inferences than do most other people.

Admittedly, some gaps in self-understanding are more surprising. For instance, Nisbett and Wilson (1977) showed that people are often wildly mistaken about what caused or led them to make even the most mundane decisions.[3] In one study, people were asked to compare the qualities of four identical pairs of nylon stockings. Many subjects judged the stockings to the right of the display to be better than those to the left—choosing the right-most over the leftmost by a factor of almost four to one—while inventing or "confabulating" grounds for their choices. All of them strongly denied that

[1] On overestimating one's own intelligence see Kruger & Dunning (1999), on physical abilities see Dunning, Meyerowitz, & Holzberg (1989), on personality traits see Messick, Bloom, Boldizar, & Samuelson (1985), and on physical attractiveness Heine & Lehman (1997).
[2] See Festinger (1957) and the work of other "dissonance" theorists.
[3] For related studies of the misperception of the role of reasons in decision-making see Shafir, Simonson, and Tversky (1993), and Russo, Meloy, and Medvec (1998).

the relative positions of the stockings played a role in their thinking.

Philosophical discussions in the modern era have not concentrated on our demonstrably imperfect knowledge of our characters and our reasons for acting, but have instead focused on the special security and immediacy of self-knowledge in its most basic incarnations. This emphasis can be traced back to Descartes' demonstration that while he could rationally doubt the existence of his brain, his body and the world around him, he could not rationally doubt the existence of his own mind. The metaphysical consequences of Descartes' epistemological argument are still hotly debated, but its perceived success has given rise to a set of nettlesome questions. First, how much about one's own mind is beyond rational doubt? Might I be mistaken in supposing that I feel pain when I am really quite comfortable? Can I falsely judge that I believe some proposition that I in fact doubt? Can I correctly ascertain that *someone* is angry and incorrectly conclude that *I* am angry by falsely assuming that I am the angry person in question? Errors of this kind are at least uncommon; are they even possible?

A second (related) set of questions surrounds the *way* in which we know basic facts about our minds. One does not typically infer that one has particular experiences, thoughts, or beliefs from premises about one's behavior. (Indeed, the immediacy of ordinary self-knowledge was a persistent source of embarrassment for philosophers and psychologists with behaviorist sympathies.) Following Locke, contemporary empiricists such as David Armstrong (1963) and (1968) and Alvin Goldman (1993) and (2006) have argued for an "inner sense" view according to which our knowledge of our own minds is relevantly like the non-inferential perceptual knowledge we have of the material objects around us. In contrast, philosophers with rationalist sympathies, such as Tyler Burge

(1996) and Sydney Shoemaker (1996), while abandoning many of Descartes' claims about the differences between knowledge of oneself and knowledge of extra-mental reality, have persisted in arguing for deep asymmetries between the two. For instance, Burge argues that the warrant someone has for believing something about her own mind is different in kind from her reasons for believing things about the world around her: whereas perceptual warrant is provided by our visual, tactile and auditory *experiences*, introspective warrant has a *conceptual* origin in our ability to reason in a critical fashion. Many of Shoemaker's arguments are aimed at establishing the metaphysical analogue of this difference. Whereas our perceptual judgments and the perceivable facts that make them true are entirely *distinct* from one another, our introspective judgments and the psychological facts that make them true enjoy a *constitutive* or mereological connection.

My primary aim in what follows is to explain and assess several of the more important skirmishes in this battle over the source and extent of basic self-knowledge. In the course of the discussion I will also try to argue for a compromise between the competing camps. Whereas Descartes and his rationalist heirs have a largely correct view of our knowledge of our own thoughts and beliefs, Locke and the contemporary empiricists offer a superior account of our knowledge of our own sensations and experiences. The truth is in the center.

**II. Thoughts: Self-Verification**

In his second meditation Descartes shows that he cannot rationally doubt that he is thinking. Of course, it is not necessarily true that Descartes is thinking. Indeed, if the

soul does not survive the death of the body, Descartes no longer exists; and if he no longer exists, he is certainly not thinking. Nevertheless, Descartes correctly reasons, to doubt something *is* to engage in a kind of thought. Thus, my doubt that I am thinking could not be sustained or borne out by the facts, and I am in a position to know this in advance. In consequence, though an evil genius of supreme power might convince me that humans have populated the Earth for millions of years when they have not, that gravity exists when it does not, or even that I have hands when I do not, he could not deceive me into thinking that I am thinking when I am not. At least some of our self-knowledge is different in kind from our knowledge of the "external" world.

Of course, an ordinary person's reasons for believing that she is thinking won't involve a rehearsal of Descartes' argument. The ordinary Joe doesn't believe that he is thinking because he has concluded that even a supremely powerful evil demon couldn't possibly mislead him on this matter. Instead, a person's *reason* for thinking that he is thinking is just the fact that he is thinking.[4] We have direct, non-inferential, non-perceptual access to facts of this kind. What Descartes shows is that we, as theorists, can see why people are proceeding in a rational way when they assume (and so tacitly believe) that they are thinking. The assumption is a good one to make because it is conceptually impossible that it be erroneous.

Following Burge (1988) we can say that my belief that I believe something is *self-verifying*: its very existence insures its truth. Not only was Descartes right in insisting on the existence of self-verifying beliefs, he was also right in drawing an important

[4] I will have more to say below about the concept of a *reason* at play in this observation, when I discuss what reasons, if any, an ordinary person will have for believing that she believes a particular proposition.

metaphysical implication from their existence: Since it is conceptually impossible that one be mistaken in believing that one is thinking, it is similarly impossible that one be mistaken in believing that one *exists*—for surely one could not be thinking if one did not exist. "I am, I exist, is necessarily true each time that I pronounce it, or that I mentally conceive it." (1641/1973, p.150) *Cogito ergo sum*.

Still, though Descartes' meditations are extraordinarily insightful, we must be careful not to overestimate their importance. How much of our knowledge of our own minds can be accurately construed as self-verifying? Because he had a rather expansive conception of thought, Descartes hypothesized a rather large class of facts about our own minds that we cannot rationally doubt:

> But it will be said that these phenomena are false and that I am dreaming. Let it be so; still it is at least quite certain that it seems to me that I see light, that I hear noise and that I feel heat. That cannot be false; properly speaking it is what is in me called feeling; and used in this precise sense that is no other thing than thinking. (1973, p. 153)

Nevertheless, two problems beset any attempt to use self-verification to argue that none of these facts about one's mind can be rationally doubted.[5] First, even if we admit a sense of 'thought' according to which all mental phenomena are thoughts, we cannot assume that one's judgments about which *mode* of thinking one is engaged in are beyond

[5] This isn't to deny that Descartes might use arguments that do not appeal to self-verification to show why we cannot rationally doubt the rather large class of facts about our own minds that he considered indubitable. (That noted, I will go on to cast some doubts of my own on this aspect of the Cartesian project in section V below.)

rational doubt. For instance, since believing is neither a kind of hearing nor a kind of seeming to hear, my judgment that I hear or seem to hear something is not self-verifying. Second, even if I can't be wrong in the assumption that I am thinking something or other, I don't typically restrict myself to beliefs of this kind. I also have firm beliefs about which *propositions* I am considering or mulling over. Descartes does not explain why I could not be mistaken in believing that I am thinking about what my uncle Billy once said when I am really thinking about my uncle Mo's banter instead.

As we will see, the first of these problems decisively undermines any attempt to show that all basic self-knowledge consists in self-verifying judgments. (To anticipate: neither my judgment that I am in pain nor my belief that I seem to see eleven lights is truly self-verifying.) But the second problem for the Cartesian project can, I think, be solved. It is, in fact, conceptually impossible that I be mistaken in believing that I am thinking about my uncle Billy. As Burge claims, "One is thinking that p in the very event of thinking knowledgeably that one is thinking it. It is thought and thought about in the same mental act" (1988, p. 116).

**III. Moore's Paradox**

Let us then move on to consider the first of the two problems with extending Descartes' reasoning to account for the whole of self-knowledge. Perhaps I can't be mistaken in believing that I bear *some* attitude toward the proposition that Billy is funny as I must entertain or grasp this proposition in order to believe that I bear some attitude towards it.

But might I be mistaken in judging that I *believe* the proposition in question when I really, in fact, either *doubt* or *disbelieve* it?

Considerations arising from Moore's Paradox (to be described below) suggest that while my belief that I believe that p might not be self-verifying, it is nevertheless *conceptually connected* to my believing that p in a substantial if somewhat weaker way. If this is right, and our perceptual beliefs are not conceptually connected to the same extent to the perceivable facts that make them true, our knowledge of our own beliefs cannot be accurately modeled as the deliverances of an inner sense.[6]

The strongest arguments for a conceptual connection between a subject's belief that she believes that p and her belief that p involve certain varieties of Moore's Paradox.[7] The proposition that a speaker would assert were she to utter, "P, but I do not believe that p," is unexceptional—indeed, this proposition is true whenever the speaker denoted by 'I' in the context of utterance is ignorant of the truth of p. But there is nevertheless something odd about *believing* this proposition. If a subject were to believe a Moore-Paradoxical proposition—if she were to believe the proposition she would express by uttering, "P, but I do not believe that p"— her belief in that proposition would

---

[6] I include the "same extent" proviso to accommodate Burge's account of perception as elaborated in his (1996) and (2003). According to Burge, though *particular* perceptual judgments are not conceptually connected to the perceivable facts that make them true, these judgments when considered *as a whole* are connected to their truth-makers because the contents of our perceptual beliefs are determined by their relation to the (perceivable) environment. Note that this makes Burge's argumentative task harder than Shoemaker's. Because Shoemaker thinks wholesale perceptual error is possible, to impugn the perceptual model of introspection he need only show that wholesale introspective error is impossible; in contrast, Burge's externalist views of perception force him to show that a rational, conceptually equipped agent can't make the same kind of single case errors that plague our perceptual judgments. See Bilgrami (2006) for discussion.

[7] G. E. Moore is supposed to have raised the "paradox" in a lecture. For its use in the current context see the amended version of Shoemaker (1995) which appears in Shoemaker (1996, pp.74-93), and Zimmerman (2006).

be *self-falsifying*. To believe the proposition expressed by "P, but I do not believe that p," I must believe both of its conjuncts. My belief in the second conjunct is true if and only if I do not believe that p. But if I believe the first conjunct I therein believe that p. So by believing the first conjunct of a Moore-paradoxical proposition I make my belief in its second conjunct false. It is therefore logically impossible for the belief I would express by uttering "P, but I do not believe that p," to be true. Believing that p but that I do not believe that p is akin to believing that I have no beliefs.[8]

So a subject's belief in a Moore-paradoxical proposition would be self-falsifying. This supports Gareth Evans' (1982, pp. 225-6) claim that grasping the rules for the use of 'belief' involves (among other things) adopting a willingness to sincerely assert the proposition one would express by uttering "I believe that p," whenever one sincerely asserts that p. Suppose one sincerely asserts that the Giants will win the Superbowl while refusing to assert that one believes that they will and one is not agnostic about whether or not one believes that the Giants will win. In such a case one exhibits the frame of mind to which one gives voice with a sincere assertion of "The Giants will win, but I do not believe that the will." And it would seem to be a reasonable condition for possession of the concept *belief* that one not self-apply that concept in such a way that self-falsifying judgments result. But my sincerely asserting that the Giants will win either entails that I believe that will or it provides paradigmatic evidence that I have that belief. (Sincere assertion that p entails belief that p if sincere assertion *just is* the

[8] Compare with "P, but I do not know that p" which is also extremely odd to assert; see Williamson (2000). I can rationally believe both that p and that I do not know that p; I can believe that p and that I do not believe that p, but I cannot do so rationally; but it is conceptually impossible that I know both that p and that I do not know or believe that p.

expression of belief.)[9] Similarly, my sincerely asserting that I do not believe that the Giants will win either entails that I believe that I do not believe that they will win or it provides paradigmatic evidence of that higher-order belief. Coming to grasp the rules for using 'belief' (or a synonymous expression) is the most obvious way to acquire the concept *belief*. It seems then that my acquiring this concept will typically dispose me to believe that I believe that p whenever I believe that p and form some judgment about whether or not I have this belief. At the very least, my possessing the concept *belief* entails that I will display (or be disposed to display) excellent evidence that I believe that I believe that p whenever I exhibit the belief that p and form some judgment about whether or not I have it. So even if our introspective beliefs are not self-verifying, the conditions we must meet to grasp the very concept *belief* seem to ensure that our first- and second-order beliefs are not independent in the same way (or to the same degree) as are observable facts and perceptual beliefs.

An at least superficially different account of the paradox falls out of neo-expressivist theories of linguistic "first-person authority"—i.e. the fact that we almost always accept present tense self-ascriptions of 'belief' without hesitation or calls for supporting evidence. When a competent speaker assertively utters, "I believe that p," what she says is true if and only if she believes that p—i.e. she literally *asserts* nothing more than that she believes that p. But in uttering, "I believe that p," a speaker will commonly *convey* more than the minimal claim that she has asserted. In appropriate contexts, she will also suggest to her audience *that p*, while signaling to them that she is not entirely sure of the proposition she claims to believe. Suppose, for example, that

[9] See Saul Kripke's weak disquotational principle, which states, "If a normal English speaker, on reflection, sincerely assents to 'p' then he believes that p" (1996, p. 388).

when asked the time, Sarah responds with, "I think that it's three o'clock," or "I believe that it's three o'clock," or "I'm pretty sure that it's three." What she has literally said here is true just in case she believes that it is three o'clock, so this proposition is the distinctively *semantic* content of her utterance. (Supposing that it is not three o'clock, Sarah cannot be justly accused of having said something false. "I only said I *thought* it was three o'clock," she can respond. "I wasn't sure.") Still, in uttering, "I think that it's three o'clock," Sarah has *pragmatically* conveyed to her audience, however hesitantly, the proposition that it is three. Otherwise, she wouldn't be answering the inquiry about the time that was posed to her. Though the truth of Sarah's utterance does not hang on the time—but depends, instead, on her belief about the time—there are standing linguistic norms that lead all competent speakers who trust her to cautiously infer from her utterance of 'I believe that it is three o'clock' that it is indeed three o'clock. Sarah exploits these pragmatic norms in asserting something about her beliefs to answer her interlocutor's question about the time.

Let us suppose that this is right. Let us suppose, that is, that semantic norms insure that a competent speaker S will use 'I believe that p' to assert that she believes that p, but that pragmatic norms insure that with the very same linguistic act she will "hesitantly" assert that p. Then every sincere assertive utterance of 'I believe that p' will be true. As Kevin Falvey explains, "Let Φ be a mentalistic verb phrase such that the utterance 'I Φ' is expressive of the mental state Φ describes. It would seem that if such an expression is sincere, then the speaker is in the given mental state. Since this is precisely what the speaker asserts in uttering these words, the sincerity of the utterance suffices for its truth" (2000, p. 73). If the neo-expressivist is right, then in uttering "I

believe that p, but it is not the case that p," I both hesitantly assert that p and boldly assert its negation; semantic and pragmatic conventions insure that in assertively uttering a Moore's paradoxical sentence I assert an outright contradiction.[10]

Of course, it is not just odd to *utter* a Moore's paradoxical sentence; it is just as irrational for a lonely thinker to believe or judge-true the proposition such a sentence is used to assert while in the privacy of her office. Indeed, we might suppose that I somehow figure out that I have been somehow caught up in the irrational belief that the Giants will win the Superbowl and that I don't believe that they will win. (We might suppose, that is, that I know that I have a self-falsifying belief.) Because I know this belief is irrational, I will do my best to hide it from my neighbors, and try what I can to drive it from my mind. But even though I am not even *disposed* to assert the Moore's-paradoxical proposition in question, the neo-expressivist cannot deny that my belief in it is irrational. Just as rationalists must explain the oddity of Moore's-paradoxical utterances that don't reflect our beliefs, so too neo-expressivists must explain the irrationality of such beliefs when they are isolated from linguistic usage and dispositions to such. The debate between the two camps therefore comes down to whether facts about the use of 'belief' are more fundamental in explaining Moore's paradox that are facts about our deployment of the concept *belief* in thought, or whether the existence of the puzzle instead owes more to thought than language.

Still, I will continue to assume, throughout this essay, that the expressivist would overstep her bounds were she to try to *replace* the traditional problems of self-knowledge with an account of our use of psychological vocabulary. Perhaps our use of 'belief' and

[10] See Bar-On (2004) and Heal (1994) for related views.

similar terms must play a central role in explaining the distinctive way in which we form and justify beliefs about our own minds, but the phenomena to be explained are not themselves linguistic.[11]

## IV. Critical Reasoning

While these reflections on Moore's Paradox emphasize what it takes to grasp the concept *belief*, Burge's (1996) arguments against the perceptual model emphasize the nature of *critical reasoning*. According to Burge, "Critical reasoning is just reasoning in which norms of reason apply to how attitudes should be affected partly on the basis of reasoning that derives from judgments about one's attitudes" (1996, p. 102). Consider, as an example of such a norm of reasoning *normative modus ponens*.

> NMP: If you believe that p and you believe that if p then q, then you ought to believe that q or abandon your belief in at least one of these premises.

If I am to apply NMP, and not merely adopt the beliefs that it says I ought to adopt for reasons that have nothing directly to do with my grasp of NMP, I must form judgments about what I believe (however tacit). I might, for example, proceed in this way by coming to believe that q upon judging that I both believe that p and believe that if p then q and that I therefore ought to believe q or abandon one of these prior beliefs. (My first-

[11] Defending the kind of realism about the mental that informs my rejection of these strategies of replacement would take us too far afield here; but for some support see the criticism of the view Wright attributes to Wittgenstein in his (1987) and (1989) as developed in Peacocke (1999) and Zimmerman (2006).

order attitude towards q is here "affected" by the application of the "norm of reasoning" NMP, where my application of this norm, in turn, "derives from [my] judgments about [my] attitudes" towards p and the proposition that if p then q.) Thus, if we accept Burge's claim that genuinely critical reasoning necessarily involves the application of norms (and not just conformity to them), we must also accept that critical reasoning essentially involves judgments about one's own beliefs.[12]

But why must these judgments be rational ones to make, or, as Burge asks, why must we be rationally *entitled* to them?[13] And might they be rational in this way without being true and (therein) constituting self-*knowledge*? Well, suppose that I am entirely unjustified or irrational in believing that I both believe that p and believe that if p then q. Should I then conclude that I ought to believe that q? If I did draw this conclusion from

[12] Burge (1996, p. 99) claims that animals and small children are limited to "blind" reasoning that does not involve introspective judgments about their own beliefs and reasons. He even admits that adult humans often reason in this way. But he refuses to label such reasoning "critical" and he insists that critical reasoning is conceptually connected to self-knowledge in a way in which it is not connected to perceptual knowledge.

[13] Burge uses 'warrant', 'entitlement' and 'justification' in a somewhat technical way to mark distinctions not captured in ordinary usage. He uses the term 'warrant' to denote the broadest positive epistemic status of which entitlement and justification are species; and on Burge's use of 'justification', to be justified in holding a belief requires a degree of reflection and sophistication one does not need to be entitled to it. "I take the notion of epistemic warrant to be broader than the ordinary notion of justification. An individual's epistemic warrant may consist in a justification that the individual has for a belief or other epistemic act or state. But it may also be an entitlement that consists in a status of operating in an appropriate way in accord with norms of reason, even when these norms cannot be articulated by the individual who has that status" (1996, p. 93). While important, these distinctions will not occupy us here. I will continue to use 'justified' in a broad way, so that a subject is justified in believing something if it is reasonable for her to believe it. (Roughly speaking, and allowing for a degree of context sensitivity in our use of 'should', a subject is justified in believing that p when it is not the case that she should not believe that p.)

beliefs I had no good reason to hold, would I be rational in coming to believe that q on its basis? Burge thinks not.

> If one's judgments about one's attitudes or inferences were not reasonable—if one had no epistemic entitlement to them—one's reflection on one's attitudes and their interrelations could add no rational element to the reasonability of the whole process. But reflection does add a rational element to the reasonability of reasoning. (1996, p. 101)

Suppose then that I rationally but *falsely* conclude that I both believe that p and believe that if p then q. Should I then judge that I ought to believe q even though (as we might suppose) I don't believe any first-order proposition that entails q or shows it to be likely? Should I believe that q even though I have no good evidence that it is true?[14] The problem here, Burge argues, is that we have bifurcated my perspective as a reasoner in an objectionable way. As we've described the case, I don't believe that p, I don't believe that if p then q, and I don't believe anything else that suggests that q is even remotely plausible. We therefore conclude, quite naturally, that of course I shouldn't believe that

[14] Of course I could treat the proposition that I believe that p and the proposition that I believe that if p then q as evidence that q, but in many cases this would display intellectual arrogance rather than rationality. If I have no distinct evidence for some non-obvious q that I don't simply remember or seem to remember, the (purported) fact that I believe q isn't very good evidence that it is true. Non-obvious future contingents all fit into this class: e.g., the fact (or purported fact) that I believe that there is sure to be a third world war is not itself a good enough reason to believe that the war will occur. Theorists who *equate* the fact that S remembers (or seems to remember) that p with the fact that S believes that p will argue for cases in which the fact that S remembers and therein believes that p is good enough reason for S to believe that p (as we are ordinarily justified in trusting our memories). But I doubt the tenability of the equation.

q. (That concludes the first lemma. ) But we have also supposed that I believe that I believe that p, I believe that I believe that if p then q, and I believe that anyone in a situation of this sort should believe that q or revise one of her original beliefs. As I feel quite sure that I both believe that p and believe that if p then q, and (as we can imagine) I am also quite confident in my false though justified belief that I believe these things for good reasons, it seems that I would be *doxastically incontinent* in failing to believe that q: I would display irrationality in refusing to believe what I (erroneously but justifiably) think that I should believe. So, we conclude, it must be that I should believe that q. That concludes the second lemma. It cannot be that at a given time a single person or "cognitive agent" both should and should not believe a given proposition. So one of our premises must be abandoned. Burge's diagnosis is that in introducing the kind of introspective error involved here—i.e. my falsely believing that I believe that p and my falsely believing that I believe that if p then q—we are forced to "bifurcate" my mind. We have to say that one sub-personal module—i.e. the first-order believer in me—ought not believe that q. And we must say that another sub-personal module—i.e. the second-order believer in me—ought to believe that q. To introduce gross introspective error without introducing a contradiction in our description of the facts about what an agent should believe, we need to divide the mind that we're describing.

Thus, we can see why Burge quite plausibly maintains that a critical reasoner who has even a minimally unified cognitive perspective will not fall into gross introspective

error. We cannot assume that the subject is both rational and entirely mistaken about her beliefs.[15]

Considerations of a similar kind lead Shoemaker (1996) to deny the possibility of "self-blindness." It is impossible, Shoemaker argues, for a fully rational agent to have a full slate of first-order mental states, and a full grasp of psychological concepts, and yet be forced to depend entirely on observation of his own verbal and non-verbal behavior when trying to learn facts about his own mind. What sort of metaphysical account of belief best makes sense of this fact? The answer, Shoemaker argues, can be found in a particular kind of functionalist analysis. Consider, for instance, that what makes something a heart is its role in pumping blood, and that what makes something a circulatory system is its role in moving substances to and from the body's cells. Not only are these organs individuated by these functions or "causal roles," the two functions are intimately related to one another: no human could have a properly functioning circulatory system without also having a properly functioning heart. The causal role played by a properly operating human circulatory system must include the pumping of blood which is the very causal role played by a properly functioning heart. Similarly, Shoemaker, hypothesizes, the causal role that individuates a particular first-order belief will overlap the causal role essential to one's belief that one has that belief. A functionalist account of this kind is needed to explain why the evidence that a rational conceptually equipped

[15] I must admit here that my gloss on Burge's position involves a great deal that is not explicitly contained in the text and omits a great deal that is. (In particular, I have not included Burge's (1996, p. 103) explanation of why a critical reasoner's introspective judgments could not be regularly "Gettierized.") I think, however, that I have accurately captured the general form and thrust of his reasoning.

person believes that p is sufficient to attribute to him an at least tacit belief that he has that belief:

> To the extent that a subject is rational, and possessed of the relevant concepts (most importantly, the concept of belief), believing that p brings with it the cognitive dispositions that an explicit belief that one has that belief would bring, and so brings with it the at least tacit belief that one has it. (1996, p. 241)

According to Shoemaker, then, someone's having those dispositions sufficient to attribute to her the belief that p will typically also be sufficient to attribute to her the belief that she believes that p; so, the dispositions one typically has when one believes that one believes that p will *include* among their number dispositions the having of which is a sufficient condition for one's believing that p. A constitutivist view of this kind is incompatible with the possibility of brute introspective errors concerning one's beliefs. It therefore provides the ammunition the rationalist needs to reject the perceptual model of our access to such states.[16]

**V. Reasons and Justification**

If Burge and Shoemaker are right, a substantially rational, conceptually equipped subject will not believe that she believes that p when she does not believe that p. But even if we

[16] It also provides an answer to Armstrong's (1963) argument that introspective knowledge must be relevantly like perceptual knowledge because, as Hume claims, there are no necessary connections between distinct existences. On Shoemaker's view, first- and second-order beliefs are not distinct existences.

accept the neo-rationalist's arguments for this conclusion, we will not yet have an adequate epistemology of introspective belief before us. For we still have no account of *how* a subject comes to acquire beliefs about her own beliefs; nor do we have any account of what *reasons*, if any, an ordinary person will have for believing that she believes certain things and not others.

An analogy with a priori knowledge will help to bring home the point. A theorist might try to argue, in a rationalist vein, that a conceptually equipped subject could not be led by mathematical reasoning to conclude that thirteen is an even number. But even if she did show this, our theorist would not have thereby furnished us with an adequate epistemological account of our belief in this basic arithmetic fact. We would still want to know what reasons a rational subject has for believing that thirteen is odd, and how such a subject comes to ground her belief in the oddity of thirteen in her grasp of these reasons.

Of course, our simple arithmetic case is fairly straightforward. Any person with the requisite concept of *oddity* knows that odd numbers are those that cannot be divided by two without remainder. Thus, a typical subject will be led via an inference to: (iii) conclude that thirteen is odd, from: (i) her knowledge that thirteen is not evenly divisible by two, along with (ii) her knowledge that numbers that cannot be so divided are odd. In such a case, a subject's *reasons* for believing that thirteen is odd are just the contents of the states of knowledge (i) and (ii) described above—that is: (a) the fact that odd numbers are not evenly divisible by two, and (b) the fact that thirteen is not evenly divisible by two.

Speaking in full generality, we can say that a *reason* to believe that p is some fact or set of facts that either entails p or makes it sufficiently likely. Of course, when conceived of in this way, the mere existence of reasons to believe that p has little bearing on whether an individual is justified or rational in believing that proposition. For some reason r to be among an individual subject's reasons for believing that p, she must stand in a substantive cognitive relation toward r—she must somehow *grasp* r, and believe that p *because* of her grasp of r. Surely, the fact that odd numbers are not cleanly divisible by two could not have been among our subject's reasons for believing that thirteen is odd if she had no access to this fact; and she could not be properly said to believe that thirteen is odd because it is not evenly divisible by two, if her understanding of oddity were entirely inert in her forming and revising her beliefs about the status of the number thirteen. But neither of these conditions ontain in our arithmetic case, as our subject's *knowing* (a) and (b) involves her *grasp* of these reasons, and by *inferring* (iii) from (i) and (ii), she comes to believe (iii) *because* of her grasp of (a) and (b).[17]

[17] It may be helpful to clarify my use of 'reason' here. Reasons for believing p either entail p or make it likely; reasons against believing the proposition entail its falsehood or make its falsehood likely. (As work in probabilistic epistemology makes clear, when we move away from the case of entailment, the relation in question—*x's being a reason for S to believe y*—will have to be relativized to S's body of knowledge and perhaps other things as well.) Of course, 'reason' has several other uses. The term can be used to pick out the brute causes of a phenomenon as when one sets out the reasons that apples fall from trees without intending to anthropomorphize, justify or recommend the behavior of fruit. And 'reason' may also be used to denote a person's state of mind when forming her beliefs so long as it "rationalizes" the adoption of the attitude in question. (One uses 'reason' in this way when citing Frank's (false) belief that an infallible Pope has asserted that homosexuality is a sin as "the reason" Frank believes in its immorality.) But ambiguity can be avoided by clearly distinguishing one's intended use of 'reason' from the others. I will accomplish this by consistently using 'reasons' to denote *facts* (or purported facts) rather than *events* or *states* of mind.

Still, while knowing and inferring are paradigmatic ways of grasping reasons and forming beliefs on their bases, they needn't be thought of as the only ways in which reasons can be accessed and utilized. Indeed, given the strong intuition that our introspective beliefs are not generated by inference, to think of these beliefs as in any way supported by reasons we must admit other ways of grasping reasons than knowing them or allow other ways of believing propositions for reasons than inferring them from distinct propositions that entail them or lend them evidential support. Indeed, an expansion of the inferential paradigm is nearly unavoidable. For we don't want to say that an ordinary person will have *no reason* for believing that she believes a given proposition. After all, if someone had no reason for believing what she does about her own mind, wouldn't that make her introspective beliefs irrational?

According to the direct access view of self-knowledge, a typical subject will indeed have a reason for believing that she believes that p when she does. But unlike the mathematical case described above, her reason for believing that she believes a given proposition won't be a *known* proposition from which she *infers* that she has the belief in question. (A rational subject doesn't question-beggingly infer that she believes that p from the known premise that she believes that p.) Instead, just as a subject's reason for (self-verifyingly) judging that she is entertaining the thought that p simply consists in the fact that she is entertaining p, an ordinary person's reason for believing that she believes that p will just be the fact that she believes that p.

How well does this account of our reasons for introspective belief match the paradigmatic case of inferential knowledge described above? Just as the fact that odd numbers are not evenly divisible by two, and the fact that thirteen is not so divisible,

together constitute a good reason to be believe that thirteen is odd, so too the fact that S believes that p is a good reason for S to believe that she believes that p. In each case, the reason-constituting fact or set of facts entails the proposition that it is a good reason to believe. Thus, maintaining the epistemologically important points of the analogy comes down to resolving two issues: (1) whether a subject can be said to grasp the fact that she believes that p—so that it can serve as her reason for believing that she has this belief—*independently* from her forming any introspective judgments about her own mind; and (2) whether a subject's grasp of the fact that she believes that p can be *operative* in the formation and maintenance of higher-order beliefs in a way analogous to that in which one's knowledge of some premises is operative when one invests one's confidence in some conclusion that they entail.

To take up the first challenge, the direct access theorist claims that S's prior access to the fact that she believes that p will consist in that belief's playing the full functional role that Shoemaker describes—a functional role that suffices for the state's being *access conscious* (a-conscious).[18] When someone believes that p in an a- conscious manner she will be led by that belief to perform certain first-order inferences, adopt certain first-order plans, and respond to new evidence in a certain way; and, according to the direct access theory, when a person believes that p in this full-blown manner, the fact that she believes that p will be *available* to her for use in second-order reasoning. Since the conditions that make a belief a-conscious can be described without invoking higher-order states of mind, the direct-access theory provides us with a non-circular account of our access to facts about our beliefs. Of course, the theory's plausibility rests on the

[18] See Block (1997) for an account of access consciousness.

details of the functional role in question. We need to evaluate the distinction drawn between a-conscious beliefs and those that are not a-conscious to assess the plausibility of the claim that in the former set of cases but not the latter, the fact that the subject believes what she does is *available* to her should she acquire or revise her beliefs about her own beliefs.[19]

This leaves us with the second challenge: do a subject's a-conscious beliefs motivate her belief in their existence? Well, it seems that so long as the relevant metaphysical (i.e. causal or mereological) connections hold between a subject's believing that p and her believing that she has this belief there will be a clear sense in which she believes that she believes that p *because* she believes that p. She will not only have or possess good reasons for believing what she does about her own beliefs; she will hold her second-order introspective beliefs for precisely these reasons. She will therefore be properly thought of as justified or rational in forming and maintaining her higher-order beliefs.

## VI. The Phenomenology of Direct Access: Evans' Procedure

There are two main sources of dissatisfaction with the direct access account of our knowledge of our own beliefs. First, theorists have argued that it does not capture the

[19] I am prevented by considerations of space from explaining how the a-consciousness of a subject's belief that p can be seen to suffice for her having access to the fact that she believes that p (so that it can serve as her reason for believing that she believes that p). Some of the details can be found in Zimmerman (2006). I do think, however, that Burge's argument lend support to the claim, for they suggest that in describing S as either tacitly or explicitly believing some first-order proposition we therein characterize some essential aspect of her epistemic situation or point of view where this point of view is reflected in our judgments about what we should or should not believe.

phenomenology of self-knowledge. In particular, it does not explain why we experience our access to our beliefs as "transparent to" our access to non-psychological reality. Can the direct access theorist accurately capture the phenomenology of introspection? Second, theorists have argued that the rationalist view cannot account for cases of prejudice and self-deception in which subjects form mistaken but introspectively justified beliefs about what they believe. Does the account overestimate the security of our self-ascriptions? Let us consider both of these objections in turn.

First, what typically goes on when a rational person forms a belief about what she does or does not believe? Evans (1982, p. 225) accurately describes a core set of these cases when he observes that one answers the question, "Do you believe that there will be a third world war?" not by immediately attending to your own mind but by considering the possible causes of another worldwide conflict. Some of those impressed by Evans' observation have tried to use it to show that our second-order beliefs are either inferentially justified or in some other way epistemically "grounded in" something distinct from the beliefs that make them true. Perhaps, for instance, my evidence, reasons or justification for believing that I believe that there will be a third world war is really constituted by the fact that I have good reasons to believe that such a war will occur—the very reasons I consider when answering "Is WWIII likely?"[20]

It is of course true that a rational agent will not believe in the inevitability of a third world war until she has assessed the evidence for and against its occurrence. Thus,

[20] Though there are important differences among them, the following theorists are all moved by Evan's procedure to embrace what I regard to be indirect accounts of self-knowledge: Dretske (1995); Fernandez (2003) and (2005); Gallois (2004); Gordon (1996); and Moran (2001) though I am not entirely sure what Moran takes the distinctively epistemic consequences of transparency to be.

insofar as I have not made up my mind as to whether we will go to war, I must first assess the political climate, form a judgment on the matter, and only then report on my attitude toward the issue. But why not think that the evidence for war grounds my first-order belief that war will occur, and that it is the fact that I have this belief that leads me to believe that I have it? It seems that the transparency account and rationalist accounts jibe equally well with "what it is like" to deploy Evans' procedure.[21]

Nevertheless, despite their equality in this regard, there is a dilemma facing indirect accounts of self-knowledge that does not afflict the direct access alternative. The indirect account says that my justification for believing that I believe that p stems from the fact that I have (or at least seem to have) good evidence that p. Either I can have or seem to have this evidence without believing that p or I cannot. Suppose I cannot. Then in every case in which I am led by my possession of evidence for p to believe that I believe that p, I also believe that p. Why, then, must the fact that I have or seem to have evidence that p play any role in justifying me in believing that I believe that p? Why can't I instead ground my introspective belief in the very fact that makes it true? Surely, if I believe that p in *every* case in which I have good evidence that p, the indirect account

[21] There are cases in which one has already made up one's mind whether p but must now reconsider the evidence to see whether it warrants a change in view. Evans' procedure is appropriate here as well. Nevertheless, the evidence for p that one rehashes in "renewing" one's belief that p needn't play any role in making one's belief that one believes that p justified or rational. One's second-order belief that one believes that p can be *exhaustively* grounded in the fact that one believes that p even though this fact only obtains because (as a rational agent) one has recalled considerable evidence that p. (This is just an instance of the phenomenon often noted in discussions of a priori justification where (empirical) enabling conditions don't affect the source or type of one's (a priori) grounds for belief.) But there are also cases in which one has already made up one's mind that p where one knows that there is no need to renew one's justification by rehearsing the evidence in its favor. The direct access view fares far better than its competitors in accounting for "what it is like" when one knows what one believes in this last group of cases.

posits an entirely gratuitous source of introspective justification. So let us suppose, instead, that the transparency theorist grants that I can have good evidence that p without believing that p. Suppose, for example, that when I reflect on international hostilities I find strong evidence that a third world war will occur, but that I am not moved by that evidence to believe in the likelihood of a coming war. Might I then rationally but mistakenly conclude that I believe that a third world war will occur? The Moore-paradoxical quality of my frame of mind suggests otherwise. Because I have correctly judged that I have good evidence for war I will confidently assert that I believe that war is likely; but because I don't believe that war is likely, I will refuse to assert that it is. Though I do not adopt a self-falsifying belief by answering, "Most certainly, yes" to "Do you believe WWIII is likely?" and then refusing to answer "Yes" to "Is WWIII likely?" the stance revealed by this pattern of response is sufficiently odd to place both my rationality and conceptual competence into question. Thus, it seems, the indirect account either posits an entirely gratuitous source of introspective justification, or it is forced to describe as 'rational' or 'justified' a range of beliefs that are intuitively bizarre. Neither horn of the dilemma is particularly palatable.[22]

Of course, if the direct access account is right, it cannot happen that I falsely conclude that I believe that p for reasons of the same sort that lead you to the true (self-knowledge-constituting) belief that you believe that p. And this introduces a second important point of difference between perceptual knowledge and our knowledge of our own beliefs. When unbeknownst to me I am hallucinating, and it visually seems to me that there is a table in front of me, I can come to the false but justified conclusion that

[22] For a full discussion see Zimmerman (2004) and (2005).

there is a table in front of me. Moreover, it is plausible, if somewhat more controversial, that in cases of this kind I have the same sort of reason for believing that there is a table in front of me as I have in those veridical cases in which I actually perceive a table. In both instances my evidence that there is a table in front of me is constituted by the fact that it seems to me that there is a table in front of me. But nothing comparable is possible with regard to my belief that I believe that there is a table in front of me. According to the direct access theorist, there is no such thing as its seeming to someone as though she believes that p that might be distinguished from her actually believing that p.

This brings us to our second objection: Doesn't the direct account overestimate the epistemic security of our introspective judgments? Peacocke (1999) seems to think so, for on his alternative account, one's second-order introspective belief that one believes that p will indeed be grounded in a "doxastic seeming" or "quasi-judgment": a conscious, occurrent act of mind that is in all respects exactly like cases in which one judges that p except for the fact that (in contrast with cases of genuine judgment) one can quasi-judge that p without believing that p (even at the time of that judgment).[23] In a case in which one comes to know that one believes that p via introspection, one's reason for believing that one believes that p will be the fact that one has quasi-judged that p, where one's quasi-judging that p is "normally connected" to one's believing that p. Nevertheless, Peacocke insists, this normal connection is not inviolable. If one quasi-judges that p without believing that p, one will come to the mistaken belief that one

[23] It is analytic that if S judges that p at t, S believes that p at t. I have replaced Peacocke's 'judgment' with 'quasi-judgment' to provide a conceptually coherent label for the kind of mental state he means to denote.

believes p and have as one's evidence for believing that one holds that belief the very evidence that supports paradigm instances of self-knowledge.

Do these kinds of errors ever really happen? Peacocke describes the kind of case he has in mind:

> Someone may [quasi] judge that undergraduate degrees from countries other than her own are of an equal standard to her own, and excellent reasons may be operative in her assertions to that effect. All the same, it may be quite clear, in decisions she makes on hiring, or in making recommendations, that she does not really have this belief at all. In making a self-ascription of a belief on the basis of a conscious [quasi] judgment, one is relying on the holding of the normal relations between [quasi] judgment and belief which are not guaranteed to hold. (1999, pp. 242-3)

Peacocke seems to reason as follows: Given that the subject in question is being forthright in uttering, "Undergraduate degrees from foreign universities are just as good," we must at least say that it seems to her that she believes the proposition she has expressed. For if she does not really believe what she has said, her assertion, if it is to be construed as sincere (in even an attenuated sense of 'sincerity'), must at least give voice to her quasi-endorsement of that proposition and the reasons she can give in its support. Nevertheless, though it seems to our subject S as though she believes that the two degrees are equal in value, we cannot say that she really does—for her discriminatory behavior shows that she is not disposed to act and reason in the ways we think essential to believing that proposition.

An extremely attractive description of the case opens up if we say that the dispositions that are strictly necessary for belief are conditional upon the presence of *attention* and *resolution*. If we say that S believes that p only if she is so disposed that were she fully attentive and resolute she would act on the information that p, and we allow that an agent's tendency to neglect p when absent-minded or weak-willed does not necessarily imply that she fails to believe that p, we will say that S's hiring behavior only shows that she does not believe in the equality of American and English degrees *if* she is paying attention to the relevant aspects of the situation and not suffering from weakness of will.[24] Though Peacocke does not describe the case in enough detail to assess whether this condition is met, it is quite possible that S fails to consider whether or not she might be giving undue influence to the home candidate. If she does not consider this matter—if her attention is not fully "turned toward it"—her displaying discriminatory behavior is fully compatible with her possessing a non-discriminatory belief. Notice, too, that this needn't mean that Peacocke's subject doesn't also hold overly nationalistic or prejudicial beliefs—even prejudicial beliefs of which she is ignorant—though it seems unlikely that her bias would be so detailed in its content as to directly represent the inferiority of American university degrees.

This suggests another way in which we can describe the prejudicial academic without countenancing quasi-judgments or mere seeming beliefs. We might say that S believes that degrees from foreign institutions are just as good as her own and that she wants to hire the best candidate for the job, but insist that S does not believe that *her* actions jeopardize the achievement of her end because she does not realize that the

[24] I argue for this account of belief in Zimmerman (forthcoming).

foreign candidate she passes over is superior to the domestic candidate she favors. Surely when S utters, "Foreign degrees are just as good as domestic ones," she expresses a different proposition than she would were she to utter, "Foreign candidate A is better than domestic candidate B," and a still different proposition is expressed when she mutters, "Stupid Yanks," beneath her breath when turning off the news. So it is fully compatible with S's using the first and last of these propositions to guide her actions and deliberations that she fail to use the middle term. Of course, it may be obvious that what S is doing is discriminatory. It may be obvious that in the circumstances in question S ought to infer that A is better than B from her belief in the equality of the two degrees and her appreciation of the remaining evidence. If so, S's hiring practices manifest culpable ignorance stemming from a blameworthy failure to reason properly. But we must be careful to distinguish culpable ignorance from flat-out lying. While the liar does not have the belief she pretends to have, the subject who does not realize the prejudicial nature of her actions fails to draw the inferences to which her more catholic beliefs commit her.[25]

These alternative interpretations of the prejudicial academic reveal just how difficult it is to think of truly uncontroversial cases of introspectively justified but nevertheless false second-order belief. That is, it is incredibly difficult, if not impossible, to come up with a scenario in which the best description of someone's frame of mind clearly has her falsely judging that she believes something, even though her epistemic situation or "point of view" is relevantly similar to the frame of mind in which ordinary

[25] More complicated cases will involve a kind of hypocrisy not reducible to lying. For an interesting discussion of this and a range of related cases see Bar-On (2004, chapter 8) and Falvey (2000, pp. 88-91).

subjects form true (knowledge-constituting) beliefs about what they believe. When wedded to the arguments of Descartes, Burge, and Shoemaker, this inability lends considerable support to the rationalist's claim that our knowledge of our own thoughts and beliefs has a "conceptual" rather than an "empirical" source. A great deal of self-knowledge is both non-perceptual and non-inferential in nature.[26]

## VII. Sensation and Experience

Some of these asymmetries between self-knowledge and perceptual knowledge will persist even when we move on from considering our access to our own thoughts and beliefs to examine our introspective awareness of our sensations and experiences. Consider, for instance, the non-doxastic or experiential sense or use of 'seems' on which a stick in water may seem bent to me even though I know that it is straight and have no inclination at all to judge that it is bent. Experiences of this sort surely play an essential role in the acquisition of perceptual knowledge. Indeed, as I've claimed above, if we put cases of subliminal perception to the side, it is reasonable to suppose that when mature, reflective people form their perceptual judgments, they almost always assume—and so at least tacitly believe—that things really are as they experientially seem to them to be. And just as it never happens that it seems to me that I believe that p when I do not, it never experientially seems to me that I am in pain when I am not, and it could not experientially seem to me that there seems to me to be something red in front of me when there doesn't

[26] I should make it clear, however, that though I have argued (in the previous section) that the rationalist's "direct access" account has the resources to respond to its critics, the view has not gained universal acceptance; for continued resistance, see, e.g., Goldman (2006).

seem to be something red there at all. We do not have introspective experiences that stand to our introspective beliefs in the relation to which our visual experiences stand to our perceptual beliefs.

Nevertheless, in stark contrast with our beliefs and thoughts, we can quite easily describe mistaken judgments about our own experiences that don't have their source in irrationality, conceptual confusion or cognitive malfunction. Admittedly, many cases that might initially look to involve someone caught in an erroneous introspective judgment probably do not. Consider, for example, the phenomenon known as *inattention blindness*. In one particularly dramatic experiment subjects were asked to view a video tape and count how many times the people on the screen passed a basketball between them. In the middle of the recording someone dressed in a gorilla suit strolls onto the scene, thumps his chest, and then walks off. 58% of observers were so caught up in tracking the ball's movement that (when asked afterward) they sincerely denied having seen the gorilla.[27] Do these subjects have visual experiences of the gorilla while believing that they do not? Or are their experiences *gappy*, representing all and only those things to which they are paying attention? Of course, experiments have demonstrated time and again that visible aspects of our environment that we do not consciously notice affect our subsequent behavior and thinking, so information about these features must be encoded by our visual systems in some way or other. But does the encoded information reach the level of conscious visual *experience*? The answer is not entirely clear.

---

[27] Simon and Chabris (1999); for an overview of the phenomenon see Chun and Marois (2002), and go to http://hvattum.net/wp/?p=3 to perform the gorilla experiment on unsuspecting acquaintances.

Still, it would be hasty to dismiss the question as unanswerable or ill formed.[28] You would have noticed, for example, if you were among the 58% of subjects who missed the gorilla, that your experience changed once the gorilla was drawn to your attention; it would not have seemed as though you then began noticing aspects of your experience that were there all along. And this provides good (though defeasible) evidence that when you pay attention to various objects in a scene over which your eyes are regularly scanning, this changes the character and content of your visual experience of the scene and not just your beliefs about that experience's character and content. That is, when inattention blindness is brought to your attention, your experience undergoes a gestalt switch similar to that produced when you realize that the sound you are hearing is not coming from rain on the roof, but is, instead, the repetitive thump of a record player left running.[29] Subjects haven't misjudged their experiences when surveying the gorilla-ridden scene; by focusing their attention so carefully on the basketball, they have narrowly restricted the scope of their experience so as to exclude a representation of the hairy beast.[30]

Nevertheless, there are a range of cases that should not be construed in this way. Perhaps the most famous involves the speckled hen which so bedeviled sense data theorists.[31] Suppose that while looking at a hen I try to figure out how many speckles it

---

[28] I have in mind here Daniel Dennett's (1991) dismissive attitude toward debates over the threshold a cognitive event must meet to be correctly classified as conscious.
[29] For a rich discussion of this and a broad range of other cases see Peacocke (1983).
[30] Of course, if this is right, then, as Eric Schwitzgebel (manuscript) emphasizes, subjects are wrong about more theoretical aspects of their experiences. If pressed they will probably judge that they visually experienced (or "took in") everything in front of them when they did not. See, too Dennett (2001) on the surprisingly narrow range of detailed visual experience.
[31] See Chisholm (1942) who says that Ryle first presented the problem to A.J. Ayer.

(experientially) seems or looks to me to have. As I am not trying to figure out how many speckles the hen in fact has, I needn't be worried about those of its speckled parts that I cannot see. Still, despite this advantage, if there are a sufficient number of visible speckles, I cannot help either wallowing in introspective ignorance or risking introspective error. If there are in fact 239 speckles visible on the hen, then it is plausible to suppose that it will look to me as though there is a hen with exactly 239 speckles in front of me. But in most cases I will either be ignorant of this fact about my experience, or if I try to count apparent speckles, I will do so incorrectly, and, in consequence, mistakenly conclude that there looks to me to be a hen in front of me with significantly more or less speckles than 239.

As with the case of inattention blindness, a theorist might insist that my experience here is "gappy" in crucial respects. Perhaps there is no number such that the there looks to me to be a hen with precisely that number of visible speckles in front of me.[32] But there is an important difference between the two cases. Whereas my experience of the first scene changes when I am told of the man in the gorilla suit, my experience of the hen does not change when I am told the number of its speckles. When I learn that the hen has exactly 239 visible speckles, I come to recognize a stable fact about my experience that I did not know before: that it has all along seemed to me as though there is a hen with exactly 239 speckles before me.

The phenomenon is perhaps most clearly exemplified in a simpler case that Evans discusses:

[32] This is Ayer's response (1940, pp. 124-5).

> Consider a case in which a subject sees ten points of light arranged in a circle, but reports that there are eleven points of light arranged in a circle, because he has made a mistake in counting, forgetting where he began. Such a mistake can clearly occur again when the subject reuses the procedure in order to gain knowledge of his internal state: his report 'I seem to see eleven points of light arranged in a circle' is just wrong. (1982, pp. 228-229)

Evans' subject does not have an irregularly malleable visual experience that morphs to match his miscalculations. Instead, he has a somewhat inaccurate view on the character and content of a largely stable stream of experience. He seems to see ten points of light in front of him, but he mistakenly thinks that he seems to see eleven.

There are at least three features of the case worth noting: Firstly, the events it describes are fairly unexceptional. Relevantly similar cases can be described that don't involve numerical concepts, as when we're asked by a doctor to characterize bodily discomfort or asked by a cook to describe the flavors of a complex broth. Secondly, Evans' case lacks even the faintest whiff of irrationality. The subject he describes is guilty of nothing more exotic than a failure to keep track when counting. Thirdly, and most importantly, the example demonstrates that though the relation between our experiences and our introspective beliefs about them is not mediated by seeming-experiences, it is mediated by cognitive or conceptual acts like counting. Moreover, though these acts are not the primary or most fundamental events operative in the fixation of our perceptual and introspective beliefs, they surely play an essential role in both processes. The mediating role that conceptualization plays in both perception of objects

and introspection of experience accounts for an important class of common errors that has no analogue in our judgments about our own thoughts.

To make perceptual judgments about the number of lights before her, a person must have some sort of "acquaintance" with those lights. That is, she must be able to distinguish the lights in question from the rest of the visible scene so as to begin counting *them* rather than some of the other things that she might then count. How will she do this? It would be wildly implausible to suppose that she might distinguish the lights by establishing a relation of *bare* acquaintance with them. (Surely she must know at least *one* substantive fact about the lights if she is to truly fix on them in thought so as to begin her calculations.) Of course, this quite reasonable condition on acquaintance is met many times over in Evans' case, for despite her misconceptions about the number of lights, Evans' subject correctly conceives of the circle of lights as a circle, as a circle of lights, as a circle of lights in front of her (and so on). But a subject's grip on some lights that she can perceive needn't be this conceptually rich to put her in a position to form beliefs about their number. So long as she can correctly conceive of them as, in her words, "These things," she can ask herself how many of them there are, and count them up in an effort to answer her question.

Similarly, it seems, if I am in a position to make judgments about how many lights I am experiencing, I must first *consider* my experience by distinguishing it from the other things of which I am now aware. Again we find that this condition is more than met in Evans' case. For despite my ignorance of how many lights I seem to see, I know that there appear to me to be some number of lights, that these lights appear to be arranged in a circle, and so on. Still, this level of sophistication is no more necessary

here than it is in the perceptual case. I can introspectively consider my experience so as to form judgments about how many things it is an experience of even if I do not know conceptually rich facts about it. I need only know a set of "demonstrative" facts comparable to those involved in perceptual judgment.

The easiest way to explain how I might secure a demonstrative conception of my experience is to suppose that I enter into Evans' scenario in a somewhat skeptical frame of mind. So let us suppose that, while staring at the lights, I consider the fact that I could be enjoying my current experience without there being lights of any kind in front of me. ("I might be hallucinating right now," I think to myself.) I then look ahead, and, using 'this' not to pick out the glimmering lights themselves, but my experience of them, I think, "This is my current, visual experience." I needn't here think of my experience as an experience of lights arranged in a circle, or even as an experience of lights. I need only think of it as my current, visual experience. (Still, if I hadn't done this—if I had instead just looked out at the lights and thought to myself "This is my experience"—I would not have judged correctly of my current, visual experience that it is my experience. I would have instead arrived at the incorrect belief that the lights before me are an experience of mine.[33]) So long as I have this sort of *minimal grip* on my visual experience, I can form true and false beliefs about it. Just as one can individuate the lights perceptually by thinking of them as "These things," one can individuate one's

[33] Similar things might be said of "visual" and "current." I must be thinking of my experience as my current *visual* experience so that the concept I deploy when using 'this' or 'that' doesn't pick out the tactile or auditory experience I am enjoying while looking at the lights. I must be thinking of my experience as my *current* visual experience so that the demonstrative concept doesn't pick out a past visual experience that might be represented by a memory that I am enjoying while looking at the lights. (For instance, while looking at the pleasant lights before me I might reflect on the candles melting into my eighth birthday cake and correctly think to myself, "That was a horrible experience.")

experience introspectively by thinking of it as "This experience." And once one's experience of the lights is individuated in this minimal way, one can begin to count the lights one seems to see.

Notice, though, that our introspective knowledge of our beliefs and thoughts is quite different in this regard. Suppose, for instance, that I'm thinking about the fact that Billy told thousands of funny jokes in his day. How do I know that this is what I am thinking? I surely do not get an initial grip on my thought by introspectively conceiving of it as *this thought* so as to then go on and "count" how many jokes I am thinking of Billy as having told. Conceptualization processes do not mediate our self-verifying introspective judgments and the embedded first-order thoughts that make them true, but they do intervene between our introspective beliefs and the experiences and sensations we know we enjoy.

Moreover, this difference between our introspective knowledge of our thoughts and our introspective knowledge of our experiences can be traced back to a difference in the *ways* in which thoughts and experiences represent things. Whereas our first-order and second-order thoughts represent things in the same medium, our introspective judgments and our experiences utilize different media. Whether one is forming or revising a first-order thought, or, instead, the belief that one has that thought, one must use concepts and processes of conceptualization of substantially the same sort. Evans's observation illustrates this in a particularly dramatic fashion, for when I form a judgment about how many lights are in front of me I must count in exactly the same way I must count if I am to arrive at a view on how many lights seem to be in front of me. But, in contrast with out thoughts, our experiences are either: (i) *non-conceptual* in content and so represent

properties (like the number of things) without our having to deploy concepts of any kind; or, at the very least, our experiences (ii) have a *modally distinct* content in that the way in which we utilize concepts in experientially representing properties (like the quantity or number of perceivable things) is different in kind from the way in which we exercise these concepts in forming beliefs about our experiences. (If I must deploy the concept *number* at all in representing the lights experientially, I can do so without the aid of counting.) It seems, therefore, that something like a *translation process* must be deployed in erecting a bridge between experiential and doxastic representation—a bridge that need not be constructed when we form beliefs about our own thoughts and beliefs.[34] When this translation process fails, the content of an introspective belief will fail to "match" (or accurately interpret) the content of the experience from which it is derived and introspective errors will inevitably result.

The following picture emerges from these reflections: In the scenario Evans imagines I first come to believe the proposition I would express by uttering, "This is my (current, visual) experience." I then conceptualize that experience in further (though equally simple) ways, as an experience of some bright things, some lights, some lights arranged in a circle, and so on. Finally, I deploy the concept *number* by counting the lights, and arrive at the incorrect judgment that I seem to see eleven points of light arranged in a circle. Schematically, with material in brackets describing the cognitive processes that lead to belief in the propositions expressed by the lettered sentences:

[34] Goldman (2006, p. 254 and pp. 260-275) makes a similar suggestion though he thinks the "introspective code" in which introspective representation takes place is distinct from the codes utilized by all of our first-order representations and he therefore takes his account in a full-blown empiricist direction that I would not take my own.

[I apply the concepts *current*, *visual*, *experience* twice over: first, to pick out my experience demonstratively, and again in correctly classifying it as my current visual experience.]

(a) This is my (current, visual) experience.

[I correctly apply the concept *things*.]

(b) It is an experience of some things.

[I correctly apply the concept *lights*.]

(c) It is an experience of some lights.

[In applying the concept *number*, I count incorrectly.]

(d) It is an experience of eleven lights.

If this is right, my reasons for believing (d)—that I seem to see eleven lights—will include (c) along with those reasons (whatever they are) that are generated in the course of my efforts to correctly count the lights.[35] In consequence, our judgments about our sensations and experiences will differ from our judgments about our thoughts and beliefs in the most epistemologically important respect. If my analysis is cogent, only our beliefs about our thoughts and beliefs are truly direct. Perhaps our knowledge of our own experiences is not inferential, but it is nevertheless (in some important sense) *indirect* in nature.[36]

---

[35] Exactly what these facts will be is a matter of controversy. I think that the best view will include among my reasons the fact I would express by uttering, "I seem to remember beginning my counting with *this* light," while perceptually discriminating the light I counted twice. But the issues surrounding memorial representation are too complex to discuss here.

[36] But might these initial judgments (a)-(c) be mistaken? Might I incorrectly judge to be an experience what is not in fact an experience? Evans (1982, p. 229) seems to think that

## VII. Conclusion

The important differences that we have uncovered between our knowledge of our thoughts and our knowledge of our experiences may just be the tip of the iceberg. We have, for instance, said nothing here about our introspective knowledge of our intentions, preferences, and emotions—and it is an open question whether knowledge of these states can be assimilated to our knowledge of our thoughts, our experiences, or some combination of the two. There is, moreover, the question of how I conceptualize myself in securing the kind of indexical reference that enables me to form introspective judgments of all kinds, and the attendant immunity to error through misidentification that Shoemaker (inspired by Wittgenstein) brought to light. In any event, a fully adequate account of self-knowledge promises to be more complex than either our empiricist or rationalist predecessors imagined it could be. Though basic self-knowledge is fairly easy to acquire, a satisfactory account of its varieties and their inner workings is not yet within reach.[37]

---

an error of this kind would be incompatible with my fully grasping the concept *experience*, but he dismisses the importance of this range of cases. "When the subject conceptualizes his experience in terms of some very elementary concept, such as a simple colour concept like 'red', it is not easy to make sense of his making a mistake…such infallibility as there is arises because we regard it as a necessary condition for the subject to possess these simple observational concepts that he be disposed to apply them when he has certain experiences. This sort of infallibility is rather limited and uninteresting. And it is of a quite different kind from that which arises in the case of self-ascription of belief."

[37] I would like to thank Brie Gertler and an anonymous referee from *Philosophy Compass* for helpful comments.

**Cited Sources**


Armstrong, D. (1963) "Is Introspective Knowledge Incorrigible?" *The Philosophical Review*, 72 (4), pp. 417-32.

Armstrong, D. (1968) *A Materialist Theory of Mind*, London: Routledge and Keegan Paul.

Ayer, A.J. (1940) *The Foundations of Empirical Knowledge*, London: Macmillan and Co.

Bar-On, D. (2004) *Speaking My Mind*, Oxford: UP.

Bilgrami, A. (2006) *Self-Knowledge and Resentment*, Harvard: UP.

Block, N. (1997) "On a Confusion about a Function of Consciousness," in *The Nature of Consciousness: Philosophical Debates,* N. Block, O. Flanagan and G. Guzeldere (eds.), Cambridge, MA: Bradford, MIT Press, pp. 375-415.

Burge, T. (1988) "Individualism and Self-Knowledge," *Journal of Philosophy*, 85, pp. 649-663.

Burge, T. (1996) "Our Entitlement to Self-Knowledge," *Proceedings of the Aristotelian Society*, 96, pp. 91-116.

Burge, T. (2003) "Perceptual Entitlement," *Philosophy and Phenomenological Research*, 57 (3), pp. 503-548.

Chisholm, R. (1942) "The Problem of the Speckled Hen," *Mind*, 51, pp. 368-373.

Chun, M. and R. Marois (2002) "The Dark Side of Visual Attention," *Current Opinion in Neurobiology*, 12, pp. 184-189.

Dennett, D. (1991) *Consciousness Explained*, Boston, MA: Little, Brown and Co.

Dennett, D. (2001) "Surprise, Surprise," *Behavioral and Brain Sciences*, 24, p. 982.

Descartes, R. (1641/1973, 1911) *The Philosophical Works of Descartes,* Vol. 1, (trans.) E.S. Haldane and G.R.T. Ross, Cambridge: UP.

Dretske, F. (1995) *Naturalizing the Mind*, Cambridge, MA: MIT Press.

Dunning, D., J.A. Meyerowitz, and A.D. Holzberg (1989) "Ambiguity and Self-Evaluation: The Role of Idiosyncratic Trait Definitions in Self-Serving Assessments of Ability," *Journal of Personality and Social Psychology*, 57, pp. 1082-1090.

Evans, G. (1982) *Varieties of Reference*, John McDowell (ed.), Oxford: UP.

Falvey, K. (2000) "The Basis of First-Person Authority," *Philosophical Topics*, 28 (2000), pp. 69-99.

Festinger, L. (1957) *Cognitive Dissonance*, Stanford: UP.

Fernandez, J. (2003) "Privileged Access Naturalized," *The Philosophical Quarterly*, 53, pp. 352-72.

Fernandez, J. (2005) "Privileged Access Revisited," *The Philosophical Quarterly*, 55, pp. 102-5.

Gallois. A. (2004) *The World Without, the Mind Within*, Cambridge: UP.

Goldman, A. (1993) "The Psychology of Folk Psychology," *Behavioral and Brain Sciences*, 16, pp. 15-28.

Goldman, A. (2006) *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, Oxford: UP.

Gordon, R. M. (1996) "'Radical' Simulationism," in P. Caruthers and P. K. Smith (ed.), *Theories of Theories of Mind*, Cambridge: UP.

Heal, J. (1994) "Moore's Paradox: A Wittgensteinian Approach," *Mind*, 103, pp. 5-24.

Heine, S. J., and D. R. Lehman (1997) "The Cultural Construction of Self-Enhancement: An Examination of Group-Serving Biases," *Journal of Personality and Social Psychology*, 72, pp. 1268-1283.

Kripke, S. (1996) "A Puzzle About Belief," in *The Philosophy of Language*, 3rd Edition, A. P. Martinich (ed.), Oxford: UP, pp. 382-410, reprinted from *Meaning and Use*, A. Margalit (ed.), Dordrecht: D. Reidel (1979), pp. 239-83.

Kruger, J., and D. Dunning (1999) "Unskilled and Unaware of it: How difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-assessments," *Journal of Personality and Social Psychology*, 77, pp. 1121-1134.

Messick, D. M., S. Bloom, J. P. Boldizar, and C. D. Samuelson (1985) "Why We Are Fairer than Others," *Journal of Experimental Social Psychology*, 21, pp. 480 -500.

Moran, R. (2001) *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton: UP.

Nisbett, R. E. and T. D. Wilson (1977) "Telling more than We Know: Verbal Reports on Mental Processes," *Psychological Review*, 84, pp. 231-59.

Peacocke, C. (1983) *Sense and Content*, Oxford: Clarendon.

Peacocke, C. (1999) *Being Known*, Oxford: Clarendon.

Russo, J. E., M. G. Meloy, and V. H. Medvec (1998) "Predecisional Distortion of Product Information," *Journal of Marketing Research*, 35, pp. 438-452.

Ryle, G. (1949) *The Concept of Mind*, New York: Harper and Row.

Schwitzgebel, E. (manuscript) "The Unreliability of Naïve Introspection."

Shafir, E., I. Simonson and A. Tversky (1993) "Reason-Based Choice," *Cognition*, 49, pp. 11-36.

Simon, D. J. and C. F. Chabris (1999) "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events," *Perception*, 28, pp. 1059-1074.

Shoemaker, S. (1995) "Moore's Paradox and Self-Knowledge," *Philosophical Studies*, 77, pp. 211-28.

Shoemaker, S. (1996) *The First-Person Perspective and Other Essays*, Cambridge: UP.

Williamson, T. (2000) *Knowledge and Its Limits*, Oxford: UP.

Wright, C. (1987) "On Making Up One's Mind: Wittgenstein on Intention," in P.

Weingartner and G. Schurz, (eds.) *Logic, Philosophy of Science and Epistemology, Proceedings of the XIth International Wittgenstein Symposium*, Vienna: Holder-Pickler Tempsky, pp. 391-404.

Wright, C. (1989) "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention," *Journal of Philosophy*, 76, pp. 622-34.

Zimmerman, A. (2004) "Unnatural Access," *Philosophical Quarterly*, 54, pp. 435-438.

Zimmerman, A. (2005) "Putting Extrospection to Rest," *Philosophical Quarterly*, 55, pp. 658-661.

Zimmerman, A. (2006) "Basic Self-Knowledge: Answering Peacocke's Criticisms of Constitutivism," *Philosophical Studies*, 128, pp. 337-379.

Zimmerman, A. (forthcoming) "The Nature of Belief," *Journal of Consciousness Studies*.