Introspection

Eric Schwitzgebel

for the *Stanford Encyclopedia of Philosophy*

draft, January 15, 2009

You can learn about your mind in the same way you learn about others' minds – by reading psychology textbooks, by observing facial expressions (in a mirror), by examining readouts of brain activity, by noting patterns of past behavior – but it's generally thought that you can also learn about your mind in a way that no one else can, by *introspection*. But what exactly is introspection? No simple characterization is widely accepted. Introspection is a process by means on which we can learn about our own minds and no one else's; but it is not the only such process.

Introspection is a key concept in epistemology, since introspective knowledge is often thought to be particularly secure, maybe even immune to skeptical doubt. Introspective knowledge is also often held to be more immediate or direct than sensory knowledge. Both of these putative features of introspection have been cited in support of the idea that introspective knowledge can serve as a ground or foundation for other sorts of knowledge.

Introspection is also central to philosophy of mind, both as a process worth study in its own right and as a court of appeal for other claims about the mind. Claims about consciousness, emotion, free will, personal identity, thought, belief, imagery, and other

mental phenomena of widespread interest have often been thought to have introspective consequences or to be susceptible to introspective verification.

1. General Features of Introspection

1.1. Necessary Features of an Introspective Process

1.2. The Targets of Introspection

1.3. The Products of Introspection

2. Approaches to Self-Knowledge

2.1. Symmetrical Accounts (Including the Theory Theory)

2.2. Self-Detection Accounts

2.2.1. Simple Self-Detection Accounts

2.2.2. Complex Self-Detection Accounts

2.3. Self-Knowledge Without Self-Detection

2.3.1. Pure Self-Shaping Accounts

2.3.2. Self-Fulfillment Accounts

2.3.3. Expressivist Accounts

2.3.4. Transparency

3. The Accuracy of Introspection

3.1. Varieties of Privilege

3.1.1. Varieties of Perfection: Infallibility, Indubitability, Incorrigibility, and Self-Intimation

3.1.2. Weaker Guarantees

3.1.3. Privilege Without Guarantee

1. General Features of Introspection

1.1. Necessary Features of an Introspective Process

For a process to qualify as "introspective" as the term is ordinarily used in contemporary philosophy of mind, it must minimally meet the following three conditions:

(1.) [*The mentality condition*] Introspection is a process that generates, or is aimed at generating, knowledge, judgments, or beliefs about *mental* events, states, or processes, and not about events transpiring outside one's mind. In this respect, it is different from sensory processes that normally deliver information about goings on outside the individual or about the non-mental aspects of the individual's body. Of course, knowledge of one's mental life might enable or subserve conclusions about matters beyond the mind. Knowledge that one is visually experiencing redness at the center of one's visual field might ground the conclusion that there is a red object directly in front

of one (whether a philosopher regards this sort of inference as a strange case or as the way perception normally works depends on her take on indirect realism about perception; see Indirect Realism and Phenomenalism; Jackson 1977; Fumerton 2006). The border between introspective and non-introspective knowledge can begin to seem blurry with respect to bodily self-knowledge such as proprioceptive knowledge about the position of one's limbs or nociceptive knowledge about one's pains. But perhaps in principle the introspective part, pertaining to the judgment about one's mind – e.g., that one has the feeling as though one's arms were crossed or of toe-ishly located pain – can be distinguished from the non-introspective judgment that one's arms are in fact crossed or one's toe is being pinched.

(2.) [*The first-person condition*] Introspection is a process that generates, or is aimed at generating, knowledge, judgments, or beliefs about *one's own mind only* and no one else's, at least not directly. Any process that generates knowledge equally of one's own and others' minds is by that token not an introspective process. Of course, introspective self-knowledge may sometimes serve as a *basis* or stepping stone to knowledge of other minds. For example, if I learn introspectively that I am angry, I might on that basis conclude that others are angry too. If simulation theorists are right, much of our knowledge of others' minds depends on first determining what our own reactions, attitudes, or other mental states are or would be, arriving at conclusions about others on the basis of an implicit or explicit parallel between them and ourselves (Gordon 1986, 1995; Goldman 1989, 2006). However, the introspective contribution to conclusions about others arrived at in such a way is at most in the discovery of my own anger; the

extension to others is itself non-introspective. Some philosophers have contemplated peculiar or science fiction cases in which we might introspect the contents of others' minds directly – for example in telepathy or when two individuals' brains are directly wired together – but the proper interpretation of such cases is disputable and in the current state of scientific knowledge they are marginal cases at best (see Gertler 2000).

(3.) [*The temporal proximity condition*] Introspection is a process that generates knowledge, beliefs, or judgments about one's *currently ongoing* mental life only; or, alternatively (or perhaps in addition) *immediately past* (or even future) mental life, within a certain narrow temporal window (sometimes called the specious present). You may know that you were thinking about Montaigne yesterday during your morning walk, but you cannot know that fact by current introspection alone – though perhaps you can know introspectively that you currently have a vivid memory of having thought about Montaigne. Likewise, you cannot know by introspection alone that you will feel depressed if your favored candidate loses the election in November – though perhaps you can know introspectively what your current attitude is toward the election or what emotion starts to rise in you when you consider the possible outcomes. Whether the target of introspection is best thought of as one's current mental life or one's immediately past mental life may depend on one's model of introspection: On self-detection models of introspection, according to which introspection is a causal process involving the detection of a mental state (see Section 2.2 below [*insert link]), it's natural to suppose that a brief lapse of time will transpire between the occurrence of the mental state that is the introspective target and the final introspective judgment about that state, which invites

(but does not strictly imply) the idea that introspective judgments generally pertain to immediately past states. On self-shaping and constitutive models of introspection, according to which introspective judgments create or embed the very state introspected (see Sections 2.3.1 and 2.3.2 below [*insert link]), it seems more natural to think that the target of introspection is one's current mental life or perhaps even (though few philosophers explicitly go so far) the immediate future.

No major contemporary philosopher of mind will call a process "introspective" if it does not meet some version of the three conditions above, though in ordinary language the temporal proximity condition may sometimes be violated. (For example, in ordinary speech we might describe as "introspective" a process of thinking about why you abandoned a relationship last month or whether you're really as kind to your children as you think you are.) However, many philosophers of mind will resist calling a process that meets these three conditions "introspective" unless it also meets some or all of the following three conditions:

(4.) [*The directness condition*] Introspection yields judgments or knowledge about one's own current mental processes relatively *directly* or *immediately*. It's difficult to articulate exactly what directness or immediacy involves in the present context, but examples of the relevant sort of indirectness or mediation should make the import of this condition relatively clear. For example, gathering sensory information about the world and then drawing theoretical conclusions based on that information should not, according to this condition, count as introspective, even if the process meets the three conditions above.

Visually discerning perceptually that a car is twenty feet in front of you and then inferring from that fact about the external world that you are having a visual experience of a certain sort does not, by this condition, count as introspective. However, as we will see in Section 2.3.4 below [*insert link], those who embrace transparency theories of introspection may reject at least strong formulations of this condition.

(5.) [*The detection condition*] Introspection involves some sort of *attunement to* or *detection of* a *pre-existing* mental state or event. For example, a process that involved creating the state of mind that one attributes to oneself would not, according to this condition, be called introspective. If I say to myself in silent inner speech, "I am saying to myself in silent inner speech, 'haecceities of applesauce'", without any idea ahead of time how I plan to complete the embedded quotation, what I say may be true, and I may know it to be true, and I may know its truth (in some sense) directly, by a means I could not know the truth of anyone else's mentality – it may meet all the four conditions above – any yet we may resist calling such a self-attribution introspective. Self-shaping (Section 2.3.1 below [*insert link]) and expressivist (Section 2.3.3 below [*insert link]) accounts of self-knowledge emphasize the extent to which our self-knowledge often does *not* involve the detection of pre-existing states; and because something like the detection condition is implicitly or explicitly accepted by many philosophers, some philosophers (including some who endorse self-shaping and expressivist views) would regard it as inappropriate to regard such accounts of self-knowledge as accounts of *introspection* proper.

(6.) [*The effort condition*] Introspection is *not constant, effortless, and automatic*. We are not every minute of the day introspecting. Introspection involves some sort of special reflection on one's own mental life that differs from the ordinary un-self-reflective flow of thought and action characteristic of daily coping. The mind may monitor itself regularly and constantly without requiring any special act of reflection by the thinker – for example, at a non-conscious level certain parts of brain or certain functional systems may monitor the goings-on of other parts of the brain and other functional systems, and this monitoring may meet all five conditions above – but this sort of thing is not what philosophers generally have in mind when they talk of introspection. However, this condition, like the directness and detection conditions, is not universally accepted. For example, philosophers who think that conscious experience requires some sort of introspective monitoring of the mind and who think of conscious experience as a more or less constant feature of our lives may reject the effort condition (Armstrong 1968, 1999; Lycan 1996).

Though not all philosophical accounts that are put forward by their authors as accounts of "introspection" meet all of conditions 4-6, most meet at least two of those conditions. Because of difference in the importance accorded to conditions 4-6, it is not unusual for different authors with otherwise similar accounts of self-knowledge to differ in their willingness to describe their accounts as accounts of "introspection".

1.2. The Targets of Introspection

Accounts of introspection differ in what they treat as the proper *targets* of the introspective process. No major contemporary philosopher believes that all of mentality is available to be discovered by introspection. For example, the cognitive processes involved in early visual processing and in the detection of phonemes are generally held to be introspectively impenetrable and nonetheless (in some important sense) mental (Marr 1983; Fodor 1983). Many philosophers also accept the existence of unconscious beliefs or desires, in roughly the Freudian sense, that are not introspectively available (e.g., Velleman 2000; Moran 2001; Wollheim 2003; though see Lear 1998). Although in ordinary English usage we sometimes say we are "introspecting" when we reflect on our character traits, contemporary philosophers of mind generally do not believe that we can directly introspect character traits in same sense that we can introspect some of our other mental states (especially in light of research suggesting that we have poor knowledge of our traits, reviewed in Taylor and Brown 1988).

The two most commonly cited classes of introspectible mental states are *attitudes*, such as beliefs, desires, evaluations, and intentions, and *conscious experiences*, such as emotions, images, and sensory experiences. (These two groups may not be wholly, or even partially, disjoint: Depending on other aspects of her view, a philosopher may regard some or all conscious experiences as involving attitudes and/or attitudes as things that are or can be consciously experienced.) It of course does not follow from the fact (if it is a fact) that some attitudes are introspectible that all attitudes are, or from the fact that some conscious experiences are introspectible that all conscious experiences are. Some accounts of introspection focus on attitudes (e.g., Nichols and Stich 2003; Byrne 2005),

while others focus on conscious experiences (e.g., Hill 1991; Goldman 2006); and it is sometimes unclear to what extent philosophers intend their remarks about the introspection of one type of target to apply to the other type. There is no guarantee that the same mechanism or process is involved in introspecting all the different potential targets.

Generically, this article will describe the targets of introspection as *mental states*, though in some cases it may be more apt to think of the targets as processes rather than states. Also, in speaking of the targets of introspection as *targets*, no presupposition is intended of a self-detection view of introspection as opposed to a self-shaping or constitutive or expressivist view (see Section 2 below [*insert link]). The targets are simply the states self-ascribed as a consequence of the introspective process if the process works correctly, or if the introspective process fails, the states that would have been self-ascribed.

1.3. The Products of Introspection

Though philosophers have not explored the issue very thoroughly, accounts also differ regarding the *products* of introspection. Most philosophers hold that introspection yields something like beliefs or judgments about one's own mind, but others prefer to characterize the products of introspection as "thoughts", "representations", "awareness", or the like, without always clarifying the relationship between these and beliefs or judgments. For ease of exposition, this article will describe the products of the

introspective process as judgments, without meaning to beg the question against competing views.

2. Approaches to Self-Knowledge

This section will outline several approaches to self-knowledge. Not all deserve to be called introspective, but an understanding of introspection requires an appreciation of this diversity of approaches – some for the sake of the contrast they provide to introspection proper and some because it's disputable whether they should be classified as introspective. These approaches are not exclusive. Surely there is more than one process by means of which we can obtain self-knowledge. Unavoidably, some of the same territory covered here is also covered, rather differently, in the entry on self-knowledge.

2.1. Symmetrical Accounts (Including the Theory Theory)

Symmetrical accounts of self-knowledge treat the process by which we acquire knowledge of our own minds as essentially the same as the process by which we acquire knowledge of other people's minds. A simplistic version of this view is that we know both our own minds and the minds of others only by observing outward behavior. Twentieth-century behaviorist principles tended to encourage this view, but no prominent treatment of self-knowledge accepts this view in its most extreme and simple form.

Bem (1972) perhaps comes closest to a simple symmetrical view, arguing on the basis of psychological research that our knowledge of the "internal states" of both self and other derives largely from the same types of behavioral evidence and employs the same principles of inference. We notice how we behave, and then we infer the attitudes that would seem to be reflected by those behaviors – and we do so even when psychologists induce the behavior without invoking the ascribed attitude. For example, Bem cites classic research in social psychology suggesting that when induced to perform an action for a small reward, people will attribute to themselves a more positive attitude toward that action than when they are induced by a large reward (Festinger and Carlsmith 1959; see also Section 3.3.2 [*insert link]). Seeing our own actions, we attribute to ourselves the same attitudes we would attribute to others performing the same actions. We know we like Thai food because we've noticed that we sometimes drive all the way across town to get it; we know that we're happy because we see or feel ourselves smiling. Bem argues that social psychology has consistently failed to show that we have any appreciable access to private information (such as remembered feelings) that might tell against such externally-driven attributions. On Bem's view, if we are better at discerning our own motives and attitudes, it's primarily because we have observed more of our own behavior than anyone else's.

Nisbett, Wilson, and their co-authors (Nisbett and Bellows 1977; Nisbett and Wilson 1977; Nisbett and Ross 1980; Wilson 2002) similarly argue for a symmetry in our knowledge of the bases and causes of our own and other's attitudes and behavior, describing cases in which people seem to show poor knowledge of these bases and causes.

For example, people queried in a suburban shopping center about why they chose a particular pair of stockings appeared to be ignorant of the influence of position on that choice, explicitly denying that influence when it was suggested to them. People asked to rate various traits of supposed job applicants were unaware that their judgments of the applicant's flexibility were greatly influenced by having been told that the applicant had spilled coffee during the job interview (see also Section 3.3.2 [*insert link]). In such cases, Nisbett and his co-investigators found that subjects' descriptions of the causal influences on their own behavior closely mirrored the influences hypothesized by outside observers. From this finding, Nisbett infers that the same mechanism drives the first-person and third-person attributions, a mechanism that that does involve any special private access to the real causes and bases of one's attitudes and behavior.

Gopnik (1993a-b; Gopnik and Meltzoff 1994) deploys developmental psychological evidence to support a symmetrical "theory theory" of self-knowledge. She points to evidence that for a wide variety of mental states, including believing, desiring, and pretending, children develop the capacity to ascribe those states to themselves at the same age they develop the capacity to ascribe those states to others. For example, children do not seem to be able to ascribe themselves past false beliefs (after having been tricked by the experimenter) any earlier than they can ascribe false beliefs to other people. This appears to be so even when that false belief is in the very recent past, having only just been revealed to be false. According to Gopnik, this pervasive parallelism shows that we are not given direct introspective access to our beliefs, desires, pretenses, and the like. Rather, we must develop a "theory of mind" in light of which we interpret evidence

underwriting our self-attributions. The appearance of the immediate givenness of one's mental states is often, Gopnik suggests, merely an "illusion of expertise": Experts engage in all sorts of tacit theorizing they don't recognize as such – the expert chess player for whom the strength of a move seems simply visually given, the doctor who immediately intuits cancer in a patient. Since we are all experts at mental state attribution, we don't recognize the layers of theory underwriting the process.

The empirical evidence behind these views remains contentious (White 1988; Nichols and Stich 2003). Furthermore, though Bem, Nisbett, Wilson, and Gopnik all stress the parallelism between mental state attribution to oneself and others and the inferential and theoretical nature of such attributions, they all also leave some room for a kind of self-awareness different in kind from the awareness one has of others' mental lives. Thus, none endorses a *purely* symmetrical view. Bem acknowledges that the parallelism only holds "to the extent that internal cues are weak, ambiguous, or uninterpretable" (1972, p. 5). With this caveat in mind, he says that our self-knowledge is only "partially" based on external cues. Nisbett and Wilson stress that we lack access only to the "processes" or causes *underlying* our behavior and attitudes. Our attitudes themselves and our current sensations, they say, can be known with "near certainty" (1977, p. 255; though contrast Nisbett and Ross 1980, p. 200-202, which seems sympathetic to Bem's skepticism about special access even to our attitudes). Gopnik allows that we "may be well equipped to detect certain kinds of internal cognitive activity in a vague and unspecified way", and that we have "genuinely direct and special access to certain kinds of first-person evidence [which] might account for the fact that we can draw some conclusions about our own

psychological states when we are perfectly still and silent", though we can "override that evidence with great ease" (1993a, p. 11-12). Ryle (1949) similarly stresses the importance of outward behavior in the (self-)attribution of mental states while acknowledging the presence of "twinges", "thrills", "tickles", and even "silent soliloquies", which we know of in our own case and that do not appear to be detectable by observing outward behavior. However, none of these authors develops an account of this apparently more direct self-knowledge. Their theories are consequently incomplete. Regardless of the importance of behavioral evidence and general theories in driving our self-attributions, in light of the considerations that drive Bem, Nisbett, Wilson, Gopnik, and Ryle to these caveats, a fully symmetrical theory of self-knowledge is probably unsustainable.

It seems reasonable to deny that symmetrical accounts of self- and other-knowledge are accounts of "introspection" strictly construed. They violate the first-person condition on accounts of introspection (condition 2 in Section 1.1 [*insert link]) – the condition that introspection be a process by means of which we can learn about our minds only – since on symmetrical accounts of self-knowledge we use the very same processes to learn about our own minds and the minds of others. They also violate the directness condition (condition 5 in Section 1.1 [*insert link]), at least if directness or immediacy is taken to exclude the sort of theory-mediated reasoning Bem, Nisbett, Wilson, and Gopnik describe. Likewise if certain *types* of mental states (such as personality traits, unconscious motives, the bases of our decisions, early perceptual processes) can only be known by cognitive processes identical to the processes by means of which we know

about those same sorts of states in other people, those mental states are not proper targets of introspection (see, e.g., Carruthers' [forthcoming a&b] argument against the introspection of attitudes, briefly discussed in Section 3.3.2 [*insert link]).

2.2. Self-Detection Accounts

Etymologically, the term "introspection" – from the Latin "looking into" – suggests a perceptual or quasi-perceptual process. Locke writes that we have a faculty of "Perception of the Operation of our own Mind" which, "though it be not Sense, as having nothing to do with external Objects; yet it is very like it, and might properly enough be call'd internal Sense" (1690/1975, p. 105, italics suppressed). Kant (1781/1997) says we have an "inner sense" by which we learn about mental aspects of ourselves that is in important ways parallel to the "outer sense" by which we learn about outer objects.

But what does it mean to say that introspection is like perception? In what respects? As Shoemaker (1994a-c/1996) points out, in a number of respects introspection is plausibly *unlike* perception. For example, introspection does not involve a dedicated organ like the eye or ear (though as Armstrong 1968 notes, neither does bodily proprioception). Nor does introspection appear to involve a distinctive phenomenology of "introspective appearances" (a point stressed by Rosenthal 2001/2005 in arguing against the perceptual analogy). The visual experience of redness has a distinctive sensory quality or phenomenology that would be difficult or impossible to convey to a blind person; analogously for the olfactory experience of smelling a banana, the auditory experience of

hearing a pipe organ, the experience of touching something painfully hot. However, no major philosopher has defended the idea that introspection has an analogously distinctive phenomenology – some quasi-sensory phenomenology in addition to, say, the visual phenomenology of seeing red (and in addition to the "cognitive phenomenology" if any that accompanies conscious thoughts in general [Siewert 1998; Horgan and Tienson 2002; Pitt 2004; Robinson 2005]) that is the phenomenology of the *introspective appearance* of the visual phenomenology of seeing red.

Contemporary proponents of quasi-perceptual models of introspection concede the existence of such disanalogies (Lycan 1996). However, we might consider an account of introspection to be quasi-perceptual, or less contentiously to be a "self-detection" account, if it meets the first five conditions described in Section 1.1 [*insert link] – that is, the mentality condition, the first-person condition, the temporal proximity condition, the directness condition, and the detection condition – and if, in addition, the following condition is met (following Shoemaker's 1994a-c/1996 characterization of the "broad perceptual" model):

(7.) [*The independent existence condition*] The target mental state and the introspective judgment *exist independently of each other* in the sense that the introspective judgment and the target mental state are distinct entities in principle ontologically separable – *causally* connected (assuming that all has gone well) but not *constitutively* connected.

2.2.1. Simple Self-Detection Accounts

Armstrong (1968, 1981, 1999) is perhaps the leading defender of a quasi-perceptual, self-detection account of introspection (see also Lycan 1996 for a similar view). He describes introspection as a "self-scanning process in the brain" (1968, p. 324), and he stresses what he sees as the important ontological distinction between the state of awareness produced by the self-scanning procedure and the target mental state of which one is aware by means of that scanning – the distinction, for example, between one's pain and one's introspective awareness *of* that pain. In stressing this distinction, he is stressing the independent existence condition described in Section 2.2. above [*insert link] and distinguishing his account from self-fulfillment or partial self-fulfillment accounts to be discussed in Section 2.3.2 below [*insert link].

Armstrong also appears to hold that the quasi-perceptual introspective process proceeds at a fairly low level cognitively – quick and simple, typically without much interference by or influence from other cognitive or sensory processes. He describes introspection as "completely non-inferential", similar to the simple detection of pressure on one's back (1968, p. 97), and he says it can be (and presumably typically is) continuous and "reflex", involving no more than keeping "a watching brief on our own current mental contents, but without making much of a deal of it" (1999, p. 115). He contrasts this reflexive self-monitoring with more sophisticated acts of deliberate introspection which he thinks are also possible (1999, p. 114). Note that in calling reflexive self-monitoring "introspection", Armstrong violates the effort condition from Section 1.1 [*insert link], which requires that introspection not be constant and automatic.

Nichols and Stich (2003) also offer a simple self-detection account. They employ a model of the mind on which having a propositional attitude such as a belief or desire is a matter of having a representation stored in a functionally-defined (and metaphorical) "belief box" or "desire box" (see also functionalism). On their account, self-awareness of these attitudes typically involves the operation of a simple "Monitoring Mechanism" that merely takes the representations from these boxes, appends an "I believe that...", "I desire that...", or whatever (as appropriate) to that representation, and adds it back into the belief box. For example, if I desire that my father flies to Hong Kong on Sunday, the Monitoring Mechanism can copy the representation in my desire box with the content "my father flies to Hong Kong on Sunday" and produce a new representation in my belief box – that is, create a new belief – with the content "I desire that my father flies to Hong Kong on Sunday". Nichols and Stich also propose an analogous but somewhat more complicated mechanism (though they leave the details unspecified) that takes percepts as its input and produces beliefs about those percepts as its output. Finally, they acknowledge the existence of a second, sometimes competing means of acquiring self-knowledge along the lines of symmetrical "theory theory" accounts like Gopnik's (see Section 2.1 above [*insert link]), though they argue that such accounts are incomplete if they do not also allow for the existence of a Monitoring Mechanism.

2.2.2. Complex Self-Detection Accounts

Goldman criticizes the account of Nichols and Stich (see Section 2.2.1 [*insert link]) for not describing how the Monitoring Mechanism detects the attitude type of the representation (belief, desire, etc.; 2006, p. 239). If talk of "belief boxes" and the like is shorthand for talk of functional role (as Nichols and Stich say), then the Monitoring Mechanism must somehow detect the functional role of the detected representation. But functional role is a matter of what is apt to cause a particular mental state and what that mental state is apt to cause (see here), and Goldman argues that a simple mechanism could not discern such dispositional and relational facts (2006, p. 247-249). Goldman also argues that the Nichols and Stich account leaves unclear how we can discern the strength or intensity of our beliefs, desires, and other propositional attitudes (2006, p. 239).

Goldman's positive account starts with the idea that introspection is a quasi-perceptual process that involves attention: "Attention seems to act like an orienting organ in introspection, analogous to the shift of eye gaze or the sniffing of the nose" (2006, p. 244). Individual attended mental states are then classified into broader categories, just as visually one might look at a ball and classify it as spherically shaped. However, on Goldman's view this process can only generate introspective knowledge of the general *types* of mental states (such as belief, happiness, bodily sensation) and some properties of those mental states (such as degree of confidence for belief, and "a multitude of finely delineated categories" for bodily sensation). Specific contents, especially of attitudes like belief, are too manifold, Goldman suggests, for pre-existing classificational categories to exist for each one. Rather, we represent the specific content of such mental states by

"redeploying" (Goldman borrows the term from Peacocke 1999) the representational content of the mental state. Redeployment involves simply copying the content of the introspected mental state into the content of the introspective belief or judgment (an aspect of Goldman's account that resembles the Nichols and Stich account, as he acknowledges). Finally, Goldman argues that some mental states require "translation" into the mental code appropriate to belief if they are to be introspected. Visual representations, he suggests, have a different format or mental code than beliefs, and therefore cognitive work will be necessary to translate the fine-grained detail of visual experience into mental contents that can be believed introspectively.

Hill (1991, forthcoming) also offers a complex quasi-sensory account of introspection. Like Goldman, Hill sees attention (in some broad, non-sensory sense) as central to introspection, though he allows for introspective awareness without attention (1991, p. 117-118). Central to Hill's account, however, is an emphasis on the capacity of introspective attention to transform – especially to amplify and enrich, even to create – the target experience. In this respect Hill argues that the introspective act differs from the paradigmatic observational act which does not transform the object perceived (though of course both scientific and ordinary – especially tactile – observation can affect what is perceived); and thus Hill's account only qualifiedly and conditionally meets the detection condition on accounts of introspection as described in Section 1.1 above [*insert link] – the condition that introspection involves attunement to or detection of a pre-existing mental state or event.

Although Armstrong, Goldman, and Hill describe their accounts as accounts of "introspection", Nichols and Stich tend not to use that term, preferring "mindreading" and "self-awareness". Indeed throughout the literature on self-knowledge, authors with otherwise similar views often differ in their readiness to use the term "introspection" to label the processes described by their accounts, complicating the question of where to draw the boundaries between accounts of introspection and accounts of non-introspective types of self-knowledge.

Self-detection accounts of self-knowledge seem to put introspection epistemically on a par with sense perception. To many philosophers, this has seemed a deficiency in these accounts. A long and widespread philosophical tradition holds that self-knowledge is epistemically special, that we have specially "privileged access" to – perhaps even infallible, indubitable, or incorrigible knowledge of – at least some portion of our mentality, in a way that is importantly different in kind from our knowledge of the world outside us. (Section 3 below [*insert link] will treat the accuracy of self-knowledge.) Both symmetrical accounts (Section 2.1 [*insert link]) and self-detection accounts (this section) of self-knowledge either deny any special epistemic privilege or characterize that privilege as similar to the privilege of being the only person to have an extended view of an object or a certain sort of sensory access to that object. The remaining accounts of self-knowledge (in Section 2.3 [*insert link]) are compatible with, and often to some extent driven by, more robust notions of the epistemic differences between self-knowledge and knowledge of environmental objects.

2.3. Self-Knowledge Without Self-Detection

There are several ways to generate judgments, or at least statements about one's own current mental life – self-attributions, let's call them – that are reliably true though they do not involve the detection of a pre-existing state. Consider the following four types of case:

(A.) *Self-attributions that prompt self-shaping:* I declare that I have a mental image of a pink elephant. At the same time I make this declaration, I deliberately cause myself to form the mental image of a pink elephant. Or: A man uninitiated in romantic love declares to a prospective lover that he is the kind of person who sends flowers to his lovers. At the same time he says this, he successfully resolves to be the kind of person who sends flowers to his lovers. The self-attribution either precipitates a change or buttresses what already exists in such a way as to make the self-attribution accurate.

(B.) *Automatically self-fulfilling self-attributions:* I think to myself, "I am thinking". Or: I judge that I am making a judgment about my own mental life. Or: I say to myself in inner speech 'I am saying to myself in inner speech: 'blu-bob'". Such self-attributions are automatically self-fulfilling. Their existence conditions are a subset of their truth conditions. In these cases, unlike the cases described in (A), no change or self-maintenance is necessary to render the self-attribution true, beyond the self-attributional event itself.

(C.) *Accurate self-attribution through self-expression:* I learn to say "I'm in pain!" instead of "ow!" as an automatic, unreflective response to painful stimuli. Or: I use the self-attributive sentence "I believe Russell changed his mind about pacifism" simply as a cautious way of expressing the belief that Russell changed his mind about pacifism, this expression being the product of reflecting upon Russell rather than a product of reflection upon my own mind. Self-expressions of this sort are assumed here to flow naturally from the states expressed in roughly the same way that facial expressions and non-self-attributive verbal expressions flow naturally from those same states – that is, without being preceded by any attempt to detect the state self-attributed.

(D.) *Self-attributions derived from judgments about the outside world:* From the non-self-attributive fact that Stanford is south of Berkeley I derive the self-attributive conclusion that I *believe* that Stanford is south of Berkeley. Or: From the non-self-attributive fact that it would be good to go to home now, I derive the self-attributive judgment that I want to go home now. These derivations may be inferences, but if so, such inferences require no specific premises about ongoing mental states. Perhaps one embraces a general inference principle like "from 'P', it is permissible to derive 'I believe that P'". Alternatively, such cases may involve a non-inferential mechanism that (sometimes) converts my beliefs that P into judgments that I believe that P.

The following accounts of self-knowledge all take advantage of one or more of these facts about self-attribution. Philosophers discussing self-knowledge often write as if approaches highlighting one of these facts conflict with approaches that highlight other of

these facts, and also as if approaches of this general sort conflict with self-detection approaches (Section 2.2. [*insert link]). While conflicts will certainly exist between different accounts intended to serve as *exhaustive* approaches to self-knowledge, it is implausible that any one or even any few of these approaches to self-knowledge is exhaustive. All four of the general approaches described above can lead, at least occasionally, to accurate self-attributions. Advocates of self-detection approaches to introspection needn't deny the existence of these other methods of arriving at accurate self-attributions. Nor do enthusiasts of any of these methods necessarily have to deny the possibility of something like detection in at least some cases – especially since the methods above seem poorly equipped, by themselves at least, to explain our self-knowledge of our own immediately past mental states. It is also hard to deny that we at least sometimes reach conclusions about our mental lives based on the kind of theoretical inference or self-interpretation emphasized by advocates of symmetrical accounts (see Section 2.1 above [*insert link]). Given all this, a broad pluralism about the sources of self-knowledge seems the only sensible course (see also Prinz 2004). There remains, of course, the empirical question of how often we employ methods of the various types – and also theoretical questions about the potential range of targets of the various methods and how exactly they work, as well as the theoretical-cum-terminological question of what methods merit the label "introspection".

2.3.1. Pure Self-Shaping Accounts

It is difficult to find accounts of self-knowledge that stress the self-shaping technique in its purest and most forward-looking form – perhaps because it's clear that self-knowledge must involve considerably more than this (Gertler forthcoming). However, McGeer (1996; McGeer and Pettit 2002) puts considerable emphasis on this aspect, writing that "we learn to use our intentional self-ascriptions to instill or reinforce tendencies and inclinations that fit with these ascriptions, even though such tendencies and inclinations may at best have been only nascent at the time we first made the judgments" (1996, p. 510). If I describe myself as brave in battle, or as a committed vegetarian – especially if I do so publicly – I create commitments and expectations for myself that help to make those self-ascriptions true. McGeer compares self-knowledge to the knowledge a driver has, as opposed to a passenger, of where the car is going: The driver, unlike the passenger, can make it the case that the car goes where she says it is going (p. 505).

There are also strains in Dennett (though Dennett may not have an entirely self-consistent view on these matters; see Schwitzgebel 2007b) that suggest a strong self-shaping view. In some places, Dennett compares self-reports about consciousness to works of fiction, immune to refutation in the same way that fictional claims are – one could no more go wrong about one's consciousness, Dennett says, than Doyle could go wrong about the color of Holmes's easy chair (e.g., 1991, p. 81, 94). This is consistent with either an anti-realist view of fiction (there are no facts about the easy chair or about consciousness; see p. 366-367) or a self-shaping realist view (Doyle *creates* facts about Holmes in making claims about him; we create facts about what it's like to be us in making claims about our consciousness, as perhaps on p. 81 and 94). More moderately, in discussing attitudes,

Dennett emphasizes how the act of formulating an attitude in language – for example, when ordering a menu item – can add a level of specification to one's attitude that was not present before, thereby partially creating the attitude self-attributed (1987, p. 20).

2.3.2. Self-Fulfillment Accounts

An emphasis on infallible knowledge through self-fulfilling self-attributions goes back at least to Augustine (c. 420 C.E./1998) and is most famously deployed by Descartes in his *Discourse on Method* (1637/1985) and *Meditations* (1641/1984), where he takes the self-fulfilling thought that he is thinking as indubitably true, immune to even the most radical skepticism, and a secure ground on which to build further knowledge.

Contemporary self-fulfillment accounts tend to exploit the idea of *containment*. In a 1988 essay, Burge writes:

> When one knows one is thinking that *p*, one is not taking one's thought (or thinking) that *p* merely as an object. 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. 654).

This is the case, Burge argues, because "by its reflexive, self-referential character, the content of the second-order [self-attributive] judgment is locked (self-referentially) onto the first-order content which it both contains and takes as its subject matter" (1988, p.

659-660; cf. Heil 1988; Gertler 2000; Heil and Gertler describe such thoughts as introspective while Burge appears not to think of self-knowledge so structured as introspective: 1998, p. 244; see also 1988, p. 652). In a 1996 essay, Burge extends his remarks to include not just "thoughts" as targets but also (certain types of) "judgments" (e.g., "I judge, herewith, that there are physical entities" and other judgments with "herewith"-like reflexivity, p. 92).

Shoemaker (1994a-c/1996) deploys the containment idea very differently, and over a much wider array of introspective targets including conscious states like pains and propositional attitudes like belief. Shoemaker speculates that the relevant containment relation holds not between the *contents or concepts employed* in the target state and in the self-ascriptive state but rather between their neural realizations in the brain. To develop this point, Shoemaker distinguishes between a mental state's "core realization" and its "total realization". One might think of mental processes as transpiring in fairly narrow regions of the brain (their core realization), and yet, Shoemaker suggests, it's not as though we could simply carve off those regions from all others and still have the mental state in question. To be the mental state it is, the process must be embedded in a larger causal network involving more of the brain (the total realization). Rrelationships of containment or overlap between core realization and total realization between the target state and the self-ascriptive judgment might then underwrite introspective accuracy. For example, the total realization of the state of pain may simply be a subset of the total realization of the state of believing that one is in pain. Introspective accuracy might then

be explained by the fact that the introspective judgment is not an independently existing state.

More recently, philosophers have applied Burge-like content-containment models (as opposed to Shoemaker-like realization containment models) to self-knowledge of conscious states or phenomenology in particular – for example, Gertler (2001), Papineau (2002), and Chalmers (2003). One possible difficulty with such accounts is that while it seems plausible to suppose that an introspective thought or judgment might contain another thought or judgment as a part, it's less clear how a self-attributive judgment or belief might contain a piece of conscious experience as a part. Beliefs, one might think, contain *concepts*, not conscious experiences, as their constituents (Fodor 1998); or, alternatively, one might think that beliefs are functional or dispositional patterns of response to input (Dennett 1987; Schwitzgebel 2002b), again rendering it unclear how a piece of phenomenology could be part of belief – though if occurrent beliefs or judgments are conscious experiences they might naturally be expected to have other conscious experiences as parts). It would seem, at least, that beliefs or judgments containing pieces of phenomenology would have to expire once the phenomenology has passed and thus that the introspective judgments could not be reinvoked or used in inferences without recreating the phenomenology. Chalmers (2003) concedes the temporal locality of such phenomenology-containing introspective judgments and consequently their limited use speech and in making generalizations. Papineau (2002), in contrast, embraces a theory in which the imaginative recreation of phenomenology in thinking about past experience is commonplace.

2.3.3. Expressivist Accounts

Wittgenstein writes:

> [H]ow does a human being learn the meaning of the names of sensations? – of the word "pain" for example. Here is one possibility: words are connected with the primitive, the natural, expressions of the sensation and used in their place. A child has hurt himself and he cries; and then adults talk to him and teach him exclamations and, later, sentences. They teach the child new pain-behaviour.
>
> "So you are saying that the word 'pain' really means crying?" – On the contrary: the verbal expression of pain replaces crying and does not describe it (1953/1968, sec. 244).

And "It can't be said of me at all (except perhaps as a joke) that I *know* I am in pain. What is it supposed to mean – except perhaps that I *am* in pain?" (1953/1968, sec. 246). On Wittgenstein's view, it is both true that I am in pain and that I say of myself that I am in pain, but the utterance in no way emerges from a process of *detecting* one's pain.

A simple expressivist view – sometimes attributed to Wittgenstein on the basis of these and related passages – denies that the expressive utterances (e.g., "that hurts!") genuinely ascribe mental states to the individuals uttering them. Such a view faces serious

difficulties accommodating the evident semantics of self-ascriptive utterances, including their use in inference and the apparent symmetries between present-tense and past-tense uses and between first-person and third-person uses (Wright 1998; Bar-On 2004). Bar-On advocates, instead, an expressivist view that preserves the similarity in logical and semantic structure between expressive and non-expressive utterances: Expressive utterances, on Bar-On's view have the same form and meaning as non-expressive utterances, though the two types of utterance are differently caused.

Expressivists have not always been entirely clear about exactly the range of target mental states expressible in this way: Vivid emotions, sudden severe pains, and at least some occurrent attitudes seem natural candidates; but is a remark about a minor and enduring discomfort in my toe similarly expressive? How about the comment that I am visually experiencing red near the center of my visual field? Is that an expression of the red visual experience, or of the belief that I am having red visual experience, or of something else, or is it not expressive in the relevant sense at all? No expressivist account has addressed the mechanisms of self-expression in sufficient detail to answer this sort of question adequately.

2.3.4. Transparency

Evans writes:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward – upon the world. If someone asks me, "Do you think there is going to be a third world war?", I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" I get myself into the position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p* (1982, p. 225).

Transparency approaches to self-knowledge, like Evans', emphasize cases in which it seems that one arrives at an accurate self-ascription not my means of attending to, or thinking about, one's own mental states, but rather by means to attending to or thinking about the external states of the world that the target mental states are about. Note that this claim has both a negative and a positive aspect: We do *not* learn about our minds by as it were gazing inward; and we *do* learn about our minds by reflecting on the aspects of the world that our mental states are about. The positive and negative theses are separable: A pluralist might accept the positive thesis without the negative one; an advocate of a symmetrical theory of self-knowledge (with respect to a certain class of target states) might accept the negative thesis without the positive. (N.B.: In the philosophical literature on self-knowledge "transparency" is also sometimes used to mean something like self-intimation in the sense of Section 3.1.1 below [*insert link], for example in Wright 1998; Bilgrami 2006. This is a completely different usage, not to be confused with the present usage.)

The range of target states to which transparency applies is a matter of some dispute. Among philosophers who accept something like transparency, belief is generally accepted as transparent (Gordon 1995, 2007; Moran 2001; Fernández 2003; Byrne 2005). Perceptual states or perceptual experiences are also often regarded as transparent in the relevant sense. Harman's example is the most cited:

> When Eloise sees a tree before her, the colors she experiences are all experienced as features of the tree and its surroundings. None of them are experienced as intrinsic features of her experience. Nor does she experience any features of anything as intrinsic features of her experiences. And that is true of you too. There is nothing special about Eloise's visual experience. When you see a tree, you do not experience any features as intrinsic features of your experience. Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree… (Harman 1990, p. 667)

Harman's emphasis here is on the negative thesis, which goes back at least to Moore (1903/1922; though Moore does not unambiguously endorse it). The view that it is impossible to attend directly to perceptual experience has recently been especially stressed by Tye (1995, 2000, 2002; see also Evans 1982; Van Gulick 1993; Shoemaker 1994a/1996; Drestke 1995; Martin 2002; Stoljar 2004).

Gordon (2007) argues (contra Nichols and Stich 2003 and Goldman 2006) that Evans-like *ascent routines* (ascending from "P" to "I believe that P") can drive the accurate self-ascription of all the attitudes, not just belief. He does this by wedding the transparency thesis to something like an expressive account of self-ascription: To answer a question about what I want – for example, which flavor ice cream do I want? – I think not about my desires but rather about the different flavors available, and then I *express* the resulting attitude self-ascriptively. Similarly for hopes, fears, wishes, intentions, regrets, etc. Gordon points out that from a very early age, before they likely have any self-ascriptive intent, children learn to express their attitudes self-ascriptively, for example with simple phrases like "[I] want banana!" (see also Bar-On 2004).

The transparency thesis by itself is consistent, not just with expressivism, but with any of the four non-detection-based self-ascription procedures described at the beginning of this section, including also self-shaping, self-fulfillment, and derivation from external facts (and indeed Aydede and Güzeldere 2005 attempt to reconcile aspects of the transparency view with a broadly detection-like approach). By itself the transparency thesis does not go far toward a positive view of the mechanisms of introspection.

Moran (2001) brings together transparency and self-shaping (and perhaps also a bit of expressivism) in his *commissive* account of self-knowledge. Moran argues that normally when we are prompted to think about what we believe, desire, or intend (and he limits his account primarily to these three mental states), we reflect on the (outward) phenomena in question and make up our minds about what *to* believe, desire, or do. Rather than

attempting to detect a pre-existing state, we open or re-open the matter and come to a resolution. Since we normally do believe, desire, and intend what we resolve to believe, desire, and do, we can therefore accurately self-ascribe those attitudes.

Byrne (2005) and Dretske (1995) bring together transparency and something like a derivational model of self-knowledge – a model on which I derive the conclusion that I believe that *P* directly from *P* itself, or (since Dretske would not accept exactly that formulation) I derive the conclusion that I am representing *x* as *F* from the fact that *x* is *F* – a fact which must of course, to serve as a premise in the derivation, be represented (or believed) by me. Byrne argues that just as one might abide by the following epistemic rule

DOORBELL: If the doorbell rings, believe that there is someone at the door

so also might someone abide by the rule

BEL: If *P*, believe that you believe that *P*.

To determine whether you believe that P, first determine whether P is the case, then follow the rule BEL.

Dretske analogizes introspection to ordinary cases of "displaced perception" – cases in which one perceives that something is the case by way of directly perceiving some other

thing (e.g., hearing that the mail carrier has arrived by hearing the dog's barking; seeing that you weigh 110 pounds by seeing the dial on the bathroom scale): One perceives that one represents *x* as *F* by way of perceiving the *F*-ness of *x*. Dretske notes, however, two points of disanalogy between the cases. In the case of hearing that the mail carrier has arrived by hearing the dog's bark, the conclusion (that the mail carrier has arrived) is only established if the premise about the dog's barking is true, and furthermore it depends on a defeasible connecting belief, that the dog's barking is a reliable indicator of the mail's arrival. In the introspective case, however, the inference, if it is an inference, does not require the truth of the premise about *x*'s being *F*. Even if *x* is not *F*, the conclusion that I'm representing x as F is supported. Nor does there seem to be any sort of defeasible connecting belief.

Tye (2000) develops a view that, like Dretske's, draws on the transparency thesis and analogizes introspection to displaced perception, though Tye limits his remarks to the introspection of conscious experience or "phenomenal character". Unlike Dretske, Tye explicitly denies that inference is involved. Instead, he proposes a mechanism similar to the sort of mechanism envisioned by simple self-detection models like those of Nichols and Stich (2003; see Section 2.2.1 [*insert link]), a reliable process that, in the case of perceptual self-awareness, takes awareness of external things as its input and yields as its output awareness of phenomenal character. The key difference between Tye's account on the one hand and the Nichols and Stich account on the other that warrants the classification of Tye's view here rather than in the section on self-detection models is this: Tye rejects the idea that the process is one of internal detection, while Nichols and Stich

stress that idea. To adjudicate this dispute (and to determine whether it might, in fact, be merely nominal), it would be helpful to have a clearer sense than has so far been given of what it means to say that one subpersonal system detects (or "monitors" or "scans") the states or contents of another.

Several authors have challenged the idea that sensory experience necessarily eludes attention – that is, they have denied the central claim of transparency theories about sensory experience. Block (1996), Kind (2003), and Smith (2008) have argued that phosphenes – those little lights you see when you press on your eyes – and visual blurriness are aspects of sensory experiences that can be directly attended. Siewert (2004) has argued that the heart of the transparency intuition is that in reflecting on sensory experience one does not *withdraw* attention from the objects sensed; but that this is compatible with devoting a certain sort of attention also to the sensory experience itself. In early discussions of attention, perceptual attention was sometimes distinguished from "intellectual attention" (James 1890/1981; Baldwin 1901-1905; see also Peacocke 1998), that is, from the kind of attention we can devote to purely imagined word puzzles or to philosophical issues. If non-sensory forms of attention are possible, then the transparency thesis for sensory experience will require restatement: Is it only *sensory* attention to sensory experience that is impossible? Or is it any kind of attention whatsoever? Simply to say we don't attend sensorily to our mental states is to make only a modest claim, akin to the claim that we see objects rather than seeing our visual experiences of objects; but to say that we cannot attend to our mental states even intellectually appears extreme. In light of this, it remains unclear how to cast the

transparency intuition to better bring out the core idea that is meant to be conveyed by the slogan that introspecting sensory experience is not a matter of attending to one's own mind.

3. The Accuracy of Introspection

3.1. Varieties of Privilege

It's plausible to suppose that people have some sort of *privileged access* to at least some of their own mental states or processes – that you know about your own mind, or at least some aspects of it, in a different way and better than you know about other people's minds, and maybe also in a different way and better than you know about the outside world. Consider pain. It seems you know your own pains differently and better than you know mine, differently and (perhaps) better than you know about the coffee cup in your hand. And if so, perhaps that special "first-person" privileged knowledge arises through something like introspection, in one or more of the senses described above in Section 2 [*insert link].

Just as there is a diversity of methods for acquiring knowledge of or reaching judgments about one's own mental states and processes, to which the label "introspection" applies with more or less or disputable accuracy, so also is there a diversity of forms of "privileged access", with different kinds of privilege and to which the idea of *access*

applies with more or less or disputable accuracy. And as one might expect, the different introspective methods do not all align equally well with the different varieties of privilege.

3.1.1. Varieties of Perfection: Infallibility, Indubitability, Incorrigibility, and Self-Intimation

A substantial philosophical tradition, going back at least to Descartes (1637/1985; 1641/1984), ascribes a kind of epistemic perfection to at least some of our judgments (or thoughts or beliefs or knowledge) about our own minds – infallibility, indubitability, incorrigibility, or self-intimation. Consider the judgment (thought, belief, etc.) that P, where P is a proposition self-ascribing a mental state or process (for example P might be *I am in pain*, or *I believe that it is snowing*, or *I am thinking of a dachshund*). The judgment that P is *infallible* just in case, if I make that judgment, it is not possible that P is false. It is *indubitable* just in case, if I make the judgment, it is not possible for me to doubt the truth of P. It is *incorrigible* just in case, if I make the judgment, it is not possible for people who know I make that judgment to be justified in believing P is false. And it is *self-intimating* it is not possible for P to be true without my reaching the judgment (thought, belief, etc.) that it is true. Note that the direction of implication for the last of these is the reverse of the first three. Infallibility, indubitability, and incorrigibility all have the form: "If I judge (think, believe, etc.) that P, then ...", while self-intimation has the form "If P, then I judge (think, believe, etc.) that P". All four theses also admit of weakening by adding conditions to the antecedent "if" clause (e.g., "If I judge that P as a result of normal introspective processes, then..."). (See Alston

1971/1989 for a helpful dissection of these distinctions. Also note that some philosophers [e.g. Armstrong 1963; Chalmers 2003] use "incorrigibility" to mean infallibility as defined here.)

Descartes (1641/1984) famously endorsed the indubitability of "I think", which he extends also to such mental states as doubting, understanding, affirming, and seeming to have sensory perceptions. He also appears to claim that the thought or affirmation that I am in such states is infallibly true. He was followed in this – especially in his infallibilism – by Locke (1690/1975), Hume (1739/1978, though see 1748/1975 for possibly a more moderate opinion), twentieth-century thinkers such as Ayer (1936/1946, 1956), Lewis (1946), and the early Shoemaker (1963), and many others. Historical arguments for indubitability and infallibility have tended to center on intuitive appeals to the apparent impossibility of doubting or going wrong about such matters as whether one is in pain or whether one is having a visual experience as of seeing red.

Recent infallibilists have added to this intuitive appeal structural arguments based on self-fulfillment accounts of self-knowledge (see Section 2.3.2. [*insert link]) – generally while also narrowing the scope of infallibility, for example to thoughts about thoughts (Burge 1988, 1996), or to "pure" phenomenal judgments about consciousness (Chalmers 2003; see also Wright 1998; Gertler 2001; Horgan, Tienson, and Graham 2006; Horgan and Kriegel 2007). The intuitive thought behind all these structural arguments is that somehow the self-ascriptive thought or judgment *contains* the mental state or process self-ascribed: the thought that I am thinking of a pink elephant contains the thought of a

pink elephant; the judgment that I am having a visual experience of redness contains the red experience itself.

In contrast, symmetrical (Section 2.1 [*insert link]) and self-detection (Section 2.2. [*insert link]) accounts of self-knowledge appear to stand in tension with infallibilism. If introspection is a causal process from mental state to an ontologically distinct self-ascription of that state, it appears that, however reliable such a process may *generally* be, there is inevitably room in principle for interference and error. Minimally, it seems, stroke, quantum accident, or clever neurosurgery could break otherwise generally reliable relationships between target mental states and the self-ascriptions of those states. Similar considerations apply to self-shaping (Section 2.3.1 [*insert link]) and expressivist (Section 2.3.3 [*insert link]) accounts, to the extent that these are interpreted casually rather than constitutively – i.e., to the extent that it is merely a causal connection between the self-ascription and the mental state self-ascribed, rather than a constitutive connection. The main idea of these two paragraphs can be summarized as follows: When the introspective or self-ascriptive judgment (or belief or whatever) somehow contains or constitutes the mental state or process introspected or self-ascribed, infallibilism appears plausible; when the connection is merely causal, infallibilism appears unsustainable.

Incorrigibility, as opposed to either infallibility or indubitability, was stressed by Rorty (1970) as "the mark of the mental" – and thus as applying to a wide range of mental states – and has also been embraced more recently by Dennett (2000, 2002). The idea behind incorrigibility, recall, is that one cannot justifiably believe in the falsity of the

self-ascriptive judgment or belief; or we might say, more qualifiedly, that if you arrive at the right kind of self-ascriptive judgment (perhaps an introspectively based judgment about a currently ongoing conscious process that survives critical reflection), then no one else, perhaps not even you in the future, knowing this, can justifiably hold that judgment to be mistaken. If I judge that right now I am in severe pain, and I do so as a result of considering introspectively whether I am indeed in such pain (as opposed to, say, merely inferring that I am in pain based on outward behavior), and if I pause to think carefully about whether I really am in pain and conclude that I indeed am, then no one else who knows this can justifiably believe that I'm not in pain, regardless of what my outward behavior might be (say, calm and relaxed) or what shows up in the course of brain imaging (say, no activation in brain centers normally associated with pain).

Incorrigibility does not imply infallibility (nor does infallibility, strictly speaking, imply incorrigibility): I may not actually be in pain, even if no one could be justified in thinking I'm not. Consequently, incorrigibility is compatible with a broader array of sources of self-knowledge than is infallibility. Neither Rorty nor Dennett, for example, appear to defend incorrigibility by appeal to self-fulfillment accounts of introspection (though in both cases, interpreting their positive accounts is difficult). Casual accounts of self-knowledge may be compatible with incorrigibility if the causal connections underwriting the incorrigible judgments are vastly more trustworthy than judgments obtained without the benefit of this sort of privileged access. Of course, unless one embraces a strict self-fulfillment account, with its attendant infallibilism, one will want to rule out abnormal

cases such as quantum accident; hence the need for qualifications (which, however, if not carefully limited, threaten to make the view vacuous or circular, like "P, unless not-P").

Self-intimating mental states are those such that, if a person (or at least a person with the right background capacities) has them, she necessarily believes or judges or knows that she does. Conscious states are often held to be in some sense self-intimating, in that the mere having of them involves, requires, or implies some sort of representation or awareness of those states. Brentano argues that consciousness, for example, of an outward stimulus like a sound, "clearly occurs together with consciousness of this consciousness", that is, the consciousness is "of the whole mental act in which the sound is presented and in which the consciousness itself exists concomitantly" (1874/1995, p. 129). Recent "higher order" and "same order" theories of consciousness (Armstrong 1968; Rosenthal 1986, 2005; Gennaro 1996; Lycan 1996; Carruthers 2005; Kriegel 2006) explain consciousness in terms of some higher order or same order thought, perception, or representation of the mental state that is conscious – the presence of this higher order or same order thought, perception, or representation being what *makes* the target state conscious. Relatedly, Horgan and others describe consciousness as "self-presenting" (Horgan, Tienson, and Graham 2005; Horgan and Kriegel 2007; the usage appears to follow Chisholm 1981, but Chisholm actually has an indubitability rather than a self-intimation thesis in mind). Shoemaker (1996) argues that beliefs – as long as they are "available" (i.e., readily deployed in inference, assent, practical reasoning, etc.), which needn't require that they are occurrently conscious – are self-intimating for individuals with sufficient cognitive capacity. Shoemaker's idea is that if the belief that P is

available in the relevant sense, then one is disposed to do things like say "I believe P", and such dispositions are themselves constitutive of believing that one believes that P.

Self-intimation claims (unlike infallibility, indubitability, and incorrigibility claims) are rarely cast as claims about "introspection". This may be because knowledge acquired through self-intimation would appear to be constant and automatic, thus violating the *effort condition* on introspection (condition 6 in Section 1.1 [*insert link]).

3.1.2. Weaker Guarantees

A number of philosophers have argued for forms of first-person privilege involving some sort of epistemic guarantee – not just conditional accuracy as a matter of empirical fact, but something more robust than that – without embracing infallibility, indubitability, incorrigibility, or self-intimation in the senses described in Section 3.1.1 above [*insert link].

Shoemaker (1968/2003), for example, argues that self-knowledge of certain psychological facts such as "I am waving my arm" or "I see a canary", when arrived at "in the ordinary way (without the aid of mirrors, etc.)", is *immune to error through misidentification relative to the first-person pronouns* (see also Campbell 1999; Pryor 1999; Bar-On 2004; Hamilton 2008). That is, although one may be wrong about waving one's arm (perhaps the nerves to one's arm were recently severed unbeknownst to you) or about seeing a canary (perhaps it's a goldfinch), one cannot be wrong due to mistakenly

identifying the person waving the arm or seeing the canary as you, when in fact it is someone else. This immunity arises, Shoemaker argues, because there is no need for identification in the first place, and thus no opportunity for *mis*-identification. In this respect, Shoemaker argues, knowledge that a particular arm that is moving is your arm (not immune to misidentification since maybe it's someone else's arm, misidentified in the mirror) is different from the knowledge that *you* are moving your arm – knowledge, that is, of what Searle (1983) calls an "intention in action".

Shoemaker has also argued for the impossibility of *self-blindness* with respect to one's beliefs, desires, and intentions, and for somewhat different reasons one's pains (1988, 1994b). A self-blind creature, on Shoemaker's conception, would be a rational creature with a conception of the relevant mental states, and who can entertain the thought that she has this or that belief, desire, intention, or pain, and who nonetheless utterly lacks introspective access to the type of mental state in question. A self-blind creature could still gain "third person" knowledge of the mental states in question, though observing her own behavior, reading textbooks, and the like. (Thus, strictly symmetrical accounts of self-knowledge of the sort described in Section 2.1 [*insert link] are accounts according to which one is self-blind in Shoemaker's sense.) Shoemaker's case against self-blindness with respect to belief turns on the dilemma of whether the self-blind creature cannot avoid "Moore-paradoxical" sentences (see Moore 1942, 1944/1993) like "it's raining but I don't believe that it's raining" in which the subject asserts both P and that she doesn't believe that P. If the subject is truly self-blind, Shoemaker suggests, there should be cases in which her best evidence is both that P and that she doesn't believe that

P (the latter, perhaps, based on misleading facts about her behavior). But if the subject asserts "P but I don't believe that P" in such cases, she does not (contra the initial supposition) really have a rational command of the nature of belief and assertion; and thus it's not a genuine case of self-blindness as originally intended. Alternatively, perhaps the creature can reliably avoid such Moore-paradoxical sentences, self-attributing belief in an apparently normal way. But then, Shoemaker suggests, it seems that she is indistinguishable from normal people in thought and behavior and hence not self-blind. For desire, intention, and pain, too, Shoemaker aims to reveal incoherences between having a rational command of the concepts in question and behaving as though one were systematically ignorant of or mistaken about those states. Shoemaker uses the case against self-blindness as part of his argument against self-detection accounts of introspection (described in Section 2.2 above [*insert link]): If introspection were a matter of detecting the presence of states that exist independently of the introspective judgment or belief, then it ought to be possible for the faculty enabling the detection to break down entirely, as in the case of blindness, deafness, etc., in outward perception (see also Nichols and Stich 2003, who argue that schizophrenia provides such a case).

Burge has influentially asserted that *brute errors* about "present, ordinary, accessible propositional attitudes [such as belief and desire]" are impossible or at least subject to "severe limits" – where a "brute error" is an error that "indicates no rational failure and no malfunction in the mistaken individual" such as commonly occur in ordinary perception due to "misleading natural conditions or look-alike substitutes" (1988, p. 657-658; 1996, p. 103-104). However, Burge offers little argument for this claim, apart from

the argument mentioned in Sections 2.3.2 and 3.1.1 above [*insert links] that for certain sorts of self-ascriptions error in general (and not just "brute error") is impossible, due to the "self-verifying" nature of such self-ascriptions.

Dretske (1995, 2004) argues that we have infallible knowledge of the *content* of our attitudes without necessarily knowing (or even having a very good idea about) the *attitude* we take toward those contents. If I believe that it will rain tomorrow, on Dretske's view, I have infallibly accurate information (which I may then access introspectively) about the fact that I am in a mental state with the content that it will rain tomorrow, but I do not have infallibly accurate information about the fact that my attitude toward that content is belief (as opposed to, say, supposition or hope or no attitude at all). This view follows from Dretske's acceptance of something like a containment account of the introspection of the *content* of the attitude (the introspective judgment employing the same content as the target attitude; see Section 2.3.2 above [*insert link], especially the discussion of Burge), while seeing knowledge of the attitude one has toward that content as requiring complex information about the causal role played by that content.

Transcendental arguments for the accuracy of certain sorts of self-knowledge offer a different sort of epistemic guarantee – "transcendental arguments" being arguments that assume the existence of some sort of experience or capacity, then develop insights about the background conditions necessary for that experience or capacity, and finally conclude that those background conditions must in fact be met. Burge (1996; see also Shoemaker 1988) argues that to be capable of "critical reasoning" one must be able to recognize

one's own attitudes, knowledgeably evaluating, identifying, and reviewing one's beliefs, desires, commitments, suppositions, etc., where these mental states are known to be the states they are. Since we are (by assumption, for the sake of transcendental argument) capable of critical reasoning, we must have some knowledge of our attitudes. Bilgrami (2006) argues that we can only be held responsible for actions if we know the beliefs and desires that "rationalize" our actions; since we can (by assumption) sometimes be held responsible, we must sometimes know our beliefs and desires. Wright (1989) argues that the "language game" of ascribing "intentional states" such as belief and desire to oneself and others requires as a background condition that self-ascriptions have special authority within that game. Given that we successfully play this language game, we must indeed have the special authority that we assume and others grant us in the context of the game.

3.1.3. Privilege Without Guarantee

Developing an analogy from Wright (1998), if it's your turn with the kaleidoscope, you have a type of privileged perspective on the shapes and colors it presents. If someone else in the room wants to know what color dominates, for example, the most straightforward course would be to ask you. But this type of privileged access comes with no guarantee. At least in principle, you might be quite wrong about the tumbling shapes. You might be dazzled by afterimages, or momentarily confused, or hallucinating, or (unbeknownst to you) colorblind. (Yes, people often don't know they are colorblind; see Kornblith 1998.) It is also at least in principle possible that others may know better than you, perhaps even systematically so, what is transpiring in the kaleidoscope. You

might think the figure shows octagonal symmetry, but the rest of us, familiar with the kaleidoscope's design, might know that the symmetry is hexagonal. A brilliant physicist may invent a kaleidoscope state detector that can dependably reveal from outside the shape, color, and position of the tumbling chunks.

Wright raises the analogy to suggest that people's privilege with respect to certain aspects of their mental lives must be different from that of the person with the kaleidoscope; but other philosophers, especially those who embrace self-detection accounts of introspection, will find the analogy at least somewhat apt: Introspective privilege is akin to the privilege of having a unique and advantageous sensory perspective on something. Metaphorically speaking, we are the only ones who can gaze directly at our attitudes or our stream of experience, while others must rely on us or on outward signs. Less metaphorically, in generating introspective judgments (or beliefs or knowledge) about one's own mentality one employs a process available to no one else. It is then an empirical question how accurate the deliverances of this process are; but on the assumption that the deliverances are in a broad range of conditions at least somewhat accurate and more accurate than the typical judgments other people make about those same aspects of your mind, you have a "privileged" perspective. Typically, advocates of self-detection models of introspection regard the mechanism or cognitive process generating introspective judgments or beliefs as highly reliable, but not infallible, and not immune to correction by other people (Armstrong 1968; Churchland 1988; Hill 1981, forthcoming; Lycan 1996; Nichols and Stich 2003; Goldman 2000, 2006).

3.2. The Danger of Self-Defeat in Concurrent Introspection

A number of authors have argued that introspection, far from being privileged, is self-defeating – that the introspective act destroys the very states or processes it attempts to assess. Comte provides an influential early statement of this view:

> But as for observing in the same way *intellectual* phenomena at the time of their actual presence, that is a manifest impossibility. The thinker cannot divide himself into two, of whom one reasons whilst the other observes him reason. The organ observed and the organ observing being, in this case, identical, how could observation take place? This pretended psychological method is then radically null and void (1830, using the translation of James 1890/1981, p. 188; see also Maudsley 1867/1977).

While Comte appears to have thought the problem fatal for scientific introspective psychology, others have embraced the point but argued that something like introspective psychology can still proceed – but through "immediate retrospection" rather than introspection, that is, through attention to or reflection on mental activity immediately *after* it transpires, while it is still in short term memory (Mill 1865/1961; James 1890/1981; Lyons 1986). Titchener (1912a&b) grants Comte's point but argues that experienced introspective observers are often capable of noticing their mental processes without disturbing them, generating results through concurrent introspection that match those obtained by immediate retrospection. Brentano (1874/1973) stresses the importance of Comte's point in undermining "inner observation" or "introspection" [*innere Beobachtung*] as opposed to "inner perception" [*innere Wahrnemung*] – the key

difference being that inner observation requires directing *attention* to the phenomena observed, which alters or destroys the phenomena (as in the case of anger, which Brentano holds to be diminished by observation), while inner perception apprehends mental processes only incidentally, without attention, and so does not disturb them.

More recently, empirical psychologists, especially Hurlburt (1990; Hurlburt and Heavey 2006; Hurlburt and Schwitzgebel 2007) and Csikszentmihalyi (Larson and Csikszentmihalyi 1983; Hektner, Schmidt, and Csikszentmihalyi 2007), have developed beeper methodologies aimed at avoiding such interference between the introspective process and the target state. Subjects in these studies wear beepers timed to sound only at long intervals, surprising them in the midst of activity and triggering an immediate retrospective assessment of their ongoing "inner experience", emotion, or thoughts.

Of course it may be true that introspection interferes with the processes introspected without thereby becoming inaccurate, especially if the introspective judgments are self-fulfilling; in such a case, one need only be wary of assuming that mental life as revealed in concurrent introspection is representative of mental life absent introspection. This issue arises particularly acutely in cases of what is sometimes called the *refrigerator light illusion*, named after the error a child might make in thinking that the refrigerator light is always on because it's always on whenever she looks to see if it is. Similarly, then, we might think we have constant visual experience of the far periphery of the visual field or constant tactile experience of our feet in our shoes, even if we do not, because whenever we think about such matters, such experience is created (Jaynes 1976; Dennett 1991;

Thomas 1999; Noë 2004; Block 2007; Schwitzgebel 2007a; among those who hold that we do have constant peripheral experiences are James 1890/1981 and Searle 1992).

3.3. Empirical Research on the Accuracy of Introspection

The arguments of the previous two sections are a priori in a broad sense of that term (the psychologists' sense): They depend on general conceptual considerations and armchair folk psychology rather than on empirical research. Now we turn to empirical research on our self-knowledge of those aspects of our minds often thought to be accessible to introspection. Since character traits are not generally thought to be introspectible aspects of our mentality, we'll skip the large literature on the accuracy or inaccuracy of our judgments about them (e.g., Taylor and Brown 1988; Funder 1999; see also Haybron's 2008 skeptical take on our knowledge of how happy we are); nor will we discuss self-knowledge of subpersonal, nonconscious mental processes, such as the processes underlying visual recognition of color and shape.

As a general matter, while a priori accounts of the epistemology of introspection have tended to stress its privilege and accuracy, empirical accounts have tended to stress its failures.

3.3.1. Self-Knowledge of the Reasons for and Causes of Attitudes and Behavior

Perhaps the most famous argument in the psychological literature on self-knowledge is Nisbett and Wilson's argument that we have remarkably poor knowledge of the reasons for and causes of our behavior and attitudes (Nisbett and Wilson 1977; Nisbett and Ross 1980; Wilson 2002). Section 2.1 above [*insert link] briefly mentioned their emblematic finding that people in a shopping mall were often ignorant of a major factor – position – influencing their judgments about the quality of pairs of stockings. In Nisbett and Bellows (1977), also briefly mentioned above, participants were asked to assess the influence of various factors on their judgments about features of a supposed job applicant. As in Nisbett and Wilson's stocking study, participants denied the influence of some factors that were in fact influential; for example, they denied that the information that they would meet the applicant influenced their judgments about the applicant's flexibility. (It actually had a major influence, as assessed by comparing the judgments of participants who were told and not told that they would meet the applicant.) Participants also attributed influence to factors that were not in fact influential; for example, they falsely reported that the information that the applicant accidentally knocked over a cup of coffee during the interview influenced "how sympathetic the person seems" to them. Nisbett and Bellows found that ordinary observers' hypothetical ratings of the influence of the various factors on the various judgments closely paralleled the participants' own ratings of the factors influencing them – a finding used by Nisbett to argue that people have no special access to causal influences on their judgments and instead rely on the same sorts of theoretical considerations outside observers rely on (the "symmetrical" view described in Section 2.1 [*insert link]). Despite some objections (such as White 1988), both psychologists and philosophers now tend to accept Nisbett's and Wilson's view that there

is at best only a modest first-person advantage in assessing the factors influencing our judgments and behavior.

In series of experiments, Gazzaniga (1995) presented commissurotomy patients (people with severed corpus callosum) with different visual stimuli to each hemisphere of the brain. With cross-hemispheric communication severely impaired due to the commissurotomy, the left hemisphere, controlling speech, had information about one part of the visual stimulus, while the right hemisphere, controlling some aspects of movement (especially the left hand) had information about a different part. Gazzaniga found that when these "split brain" patients were asked to explain why they did something, when that action was clearly caused by input to the right, non-verbal hemisphere, the left hemisphere would fluently confabulate an explanation. For example, Gazzaniga presented an instruction like "laugh" to the right hemisphere, making the patient laugh. When asked why he laughed, the patient would say something like "You guys come up and test us every month. What a way to make a living!" (p. 1393). When a chicken claw was shown to the left hemisphere and snow scene to the right, and the patient was asked to select an appropriate picture from an array, the right hand would go to a chicken and the left hand to a snow shovel; when asked why she selected those two things, the patient would say something like "Oh, that's simple. The chicken claw goes with the chicken and you need a shovel to clean out the chicken shed" (ibid.). Similar confabulation about motives is sometimes (but not always) seen in people whose behavior is, unbeknownst to them, driven by post-hypnotic suggestion (Richet 1884; Moll 1889/1911), and in

disorders such as hemineglect, hysterical blindness, and Korsakoff's syndrome (Hirstein 2005).

In a normal population, Johansson and collaborators (Johansson et al. 2005; Johansson et al. 2006) manually displayed to participants pairs of pictures of women's faces. On each trial, the participant was to point to the face he found more attractive. The picture of that face was then centered before the participant while the other face was hidden. On some trials, participants were asked to explain the reasons for their choices while continuing to look at the selected face. On a few key trials, the experimenters used slight-of-hand to present to the participant the face that was *not* selected as though it had been the face selected. Strikingly, the switch was noticed only about 28% of the time. What's more, participants actually gave explanations for their choice that appealed to specific features of the unselected face that were not possessed by the selected face 13% of the time. For example, one participant claimed to have chosen the face before him "because I love blondes" when in fact he had chosen a dark-haired face (Johansson et al. 2006, p. 690). Johansson and colleagues failed to find any systematic differences in the explanations of choice between the manipulated and non-manipulated trials, using a wide variety of measures. They found, for example, no difference in linguistic markers of confidence, emotionality, specificity of detail, complexity of description, or general position in semantic space. These results, like Nisbett's and Wilson's, suggest that at least some of the time when people think they are explaining the bases of their decisions, they are instead merely theorizing or confabulating.

Wegner has found that people can fairly readily be manipulated into believing that they willed or intended behavior that is in fact caused by another person's manipulation and, conversely, that they exerted no control over movements that were in fact their own – as with Ouija boards, with or without a cheating, intentionally directive confederate (Wegner and Wheatley 1999; Wegner 2002). The literature on "cognitive dissonance" is replete with cases in which participants' attitudes appear to change for reasons they do, or would, deny. According to cognitive dissonance theory, when people behave or appear to behave counternormatively (e.g., incompetently, foolishly, immorally), people will tend to adjust their attitudes so as to make the behavior seem less counternormative or "dissonant" (Festinger 1957; Aronson 1968; Cooper and Fazio 1984; Stone and Cooper 2001). For example, people induced to falsely describe as enjoyable a monotonous task they've just completed will tend, later, to report having a more positive attitude toward the task then those not induced to lie (though much less so if they were handsomely paid to lie in which case the behavior is not clearly counternormative; Festinger and Carlsmith 1959). Presumably, if such attitude changes were known to the subject they would generally fail to have their dissonance-reducing effect. Research psychologists have also confirmed such familiar phenomena as "sour grapes" (Lyubomirsky and Ross 1999; Kay, Jiminez, and Jost 2002) and "self-deception" (Mele 2001) which presumably also involve ignorance of the factors driving the relevant judgments and actions. And of course the Freudian psychoanalytic tradition has also long held that people often have only poor knowledge of their motives and the influences on their attitudes (Wollheim 1981; Cavell 2006).

In light of this empirical research, no major philosopher now holds (perhaps no major philosopher ever held) that we have infallible, indubitable, incorrigible, or self-intimating knowledge of the causes of our judgments, decisions, and behavior. Perhaps weaker forms of privilege also come under threat. But the question arises: Whatever failures there may be in assessing the reasons for and causes of our attitudes and behavior, are those failures failures of *introspection*, properly construed? Psychologists tend to cast these results as failures of "introspection", but if it turns out that a very different and more trustworthy process underwrites our knowledge of some other aspects of our minds – such as what our present attitudes *are* (however caused) or our currently ongoing or recently past conscious experience – then perhaps we can call only *that* process "introspection", thereby retaining some robust form of introspective privilege while acceding to the psychological consensus regarding (what we would now call non-introspective) first-person knowledge of causes and motives. Indeed, few contemporary philosophical accounts of introspection or privileged self-knowledge highlight, as the primary locus of privilege, the causes and motives of our attitudes and behavior (though Bilgrami 2006 is a notable exception).

3.3.2. Self-Knowledge of Attitudes

Research psychologists have generally not been as skeptical of our knowledge of our attitudes as they have been of our knowledge of the *causes* of our attitudes (Section 3.3.1 above [*insert link]). In fact, many of the same experiments that purport to show inaccurate knowledge of the causes of our attitudes nonetheless rely unguardedly on self-

report for assessment of the attitudes themselves – a feature of those experiments criticized by Bem (1967). For example, attitudinal surveys in psychology and social science generally rely on participants' self-report as the principal source of evidence about attitudes (de Vaus 1985/2002; Sirken et al., eds., 1999). However, as in the case of motives and causes, there's a long tradition in clinical psychology skeptical of our self-knowledge of our attitudes, giving a large role to "unconscious" motives and attitudes.

A key challenge in assessing the accuracy of people's beliefs or judgments about their attitudes is the difficulty of accurately measuring attitudes independently of self-report. There is at present no tractable measure of attitude that is generally seen by philosophers as overriding individuals' own reports about their attitudes. However, in the psychological literature, "implicit" measures of attitudes – measures of attitudes that do not reply on self-report – have recently been gaining considerable attention (see Wittenbrink and Schwarz, eds., 2007; Petty, Fazio, and Briñol, eds., 2009). Such measures are sometimes thought capable of revealing unconscious attitudes or implicit attitudes either unavailable to introspection or erroneously introspected (Wilson, Lindsey, and Schooler 2000; Kihlstrom 2004; Lane et al. 2007).

Much of the leading research on implicit attitude measures has concerned racism, in accord with the view that racist attitudes, though common, are considered socially undesirable and therefore often not self-attributed even when present. For example, Campbell, Kruskal, and Wallace (1966) explored the use of seating distance as an index of racial attitudes, noting that black and white students tended to aggregate in classroom

seating arrangements. Using facial electromyography (EMG), Vanman et al. (1997) found White participants to display facial responses indicative of negative affect more frequently when asked to imagine co-operative activity with Black than with White partners – results interpreted as indicative of racist attitudes. Cunningham et al. (2004) showed White and Black faces to White participants while participants were undergoing fMRI brain imaging. They found less amygdala activation when participants looked at faces from their own group than when participants looked at other faces; and since amygdala activation is generally associated with negative emotion, they interpreted this tendency suggesting a negative attitude toward outgroup members (see also Hart et al 1990; and for discussion Ito and Cacioppo 2007).

Much of the recent implicit attitude research has focused on response priming and interference in speeded tasks. In priming research, a stimulus (the "prime") is briefly displayed, followed by a mask that hides it, and then a second stimulus (the "target") is displayed. The participant's task is to respond as swiftly as possible to the target, typically with a classification judgment. In *evaluative priming,* for example, the participant is primed with a positively or negatively valenced word or picture (e.g., snake), then asked to make a swift judgment about whether the subsequently presented target word (e.g., "disgusting") is good or bad, or has some other feature (e.g., belongs to a particular category). Generally, negative primes will speed response for negative targets while delaying response for positive targets, and positive primes will do the reverse. Researchers have found that photographs of Black faces, whether presented visibly or so quickly as to be subliminal, tend to facilitate the categorization of negative

targets and delay the categorization of positive targets for White participants – a result widely interpreted as revealing racist attitudes (Fazio et al. 1995; Dovidio et al. 1997; Wittenbrink, Judd, and Park 1997). In the *Implicit Association Test,* respondents are asked to respond disjunctively to combined categories, giving for example one response if they see either a dark-skinned face *or* a positively valenced word and a different response if they see either a light-skinned face *or* a negatively valenced word. As in evaluative priming tasks, White respondents tend to respond more slowly when asked to pair dark-skinned faces with positively valenced words than with negatively valenced words, which is interpreted as revealing a negative attitude or association (Greenwald, McGhee, and Schwartz 1998; Lane et al. 2007).

As mentioned above, such implicit measures are often interpreted as revealing attitudes to which people have poor or no introspective access. The evidence that people lack introspective knowledge of such attitudes generally turns on the low correlations between such implicit measures of racism and more explicit measures such as self-report – though due to the recognized social undesirability of racial prejudice, it is difficult to disentangle self-presentational from self-knowledge factors in self-reports (Fazio et al. 1995; Greenwald, McGhee, and Schwartz 1998; Wilson, Lindsey, and Schooler 2000; Greenwald and Nosek 2009). People who appear racist by implicit measures may disavow racism and inhibit racist patterns of response on explicit measures (such as when asked to rate the attractiveness of faces of different races) because they don't want to be seen as racist – a motivation that may drive them whether or not they have accurate self-knowledge of their racist attitudes. Still, it seems prima facie plausible that people have

at best limited knowledge of the patterns of association that drive their responses on priming and other implicit measures.

Another question is what is revealed by participants' patterns of response on such implicit measures. In philosophy, Zimmerman (2007) and Gendler (forthcoming-a&b) have argued that measures like the Implicit Association Test do not measure actual racist *beliefs* but rather something else, under less rational control (Gendler calls them "aliefs"). In psychology, Gawronski and Bodenhausen (2006) advance a model according to which there is a substantial difference between implicit attitudes, defined in terms of associative processes, and explicit attitudes which have a propositional structure and are guided by standards of truth and consistency (see also Wilson, Lindsey, and Schooler 2000; Greenwald and Nosek 2009). On such a model, as on Zimmerman's and Gendler's views, a person with implicit racist associations may nonetheless have fully and genuinely egalitarian propositional beliefs. To the extent attitudes are held to be reflected in, or even defined by, our explicit judgments about the matter in question and also, differently but perhaps not wholly separably (see Section 2.3.4 [*insert link]), our explicit judgments about our *attitudes* toward the matter in question, our self-knowledge would seem to be correspondingly secure and implicit measures beside the point. To the extent attitudes are held to crucially involve swift and automatic, or unreflective, patterns of reaction and association, our self-knowledge of them would appear to be correspondingly problematic, corrigible by data from implicit measures.

In a different vein, Carruthers (forthcoming-a&b; see also Rosenthal 2001/2005; Bem 1967, 1972) argues that the evidence of Nisbett, Gazzaniga, Wegner, and others (reviewed in Section 3.3.1 above [*insert link]) shows that people confabulate not just in reporting the *causes* of their attitudes but also in reporting the attitudes themselves. For example, Carruthers suggests that if someone in Nisbett and Wilson's famous 1977 study confabulates "this pair seems softer" as an explanation of her choice of the rightmost pair of stockings, she errs not only about the cause of her choice but also in ascribing to herself the judgment that the pair was softer. Partly on this basis, Carruthers adopts a symmetrical view (see Section 2.1. above [*insert link]) of our self-knowledge of our attitudes while holding that we can introspect, in a stricter sense, perceptual and imagistic mental events.

3.3.3. Self-Knowledge of Conscious Experience

Currently ongoing conscious experience – or maybe immediately past conscious experience (if we hold that introspective judgment must temporally succeed the state or process introspected, or if we take seriously the concerns raised in Section 3.2 [*insert link] about the self-undermining of the introspective process) – is both the most universally acknowledged target of the introspective process and the target most commonly thought to be known with a high degree of privilege. Infallibility, indubitability, incorrigibility, and self-intimation claims (see Section 3.1.1 [*insert link]) are most commonly made for self-knowledge of states such as being in pain or having a visual experience as of the color red, where these states are construed as qualitative states,

or subjective experiences, or aspects of our phenomenology or consciousness. (All these terms are intended interchangeably to refer to what Block [1995/1997], Chalmers [1996], and other contemporary philosophers call "phenomenal consciousness".) If attitudes are sometimes conscious, then we may also be capable of introspecting those attitudes as part of our capacity to introspect conscious experience generally (Goldman 2006; Hill forthcoming).

The accuracy of self-attributions of conscious experience is difficult to study due to methodological difficulties similar to those that dog studies of implicit attitudes: There's no widely accepted measure to trump or confirm self-report. In the medical literature on pain, for example, no behavioral or physiological measure of pain is generally thought capable of overriding self-report of current pain, despite the fact that scaling issues remain a problem both within and especially between subjects (Williams, Davies, and Chadury 2000) as does retrospective assessment (Redelmeier and Kahneman 1996). When physiological markers of pain and self-report dissociate, it's by no means clear that the physiological marker should be taken as the more accurate index. (Fortunately, there appears to be, at least under tightly controlled laboratory conditions, a close relationship between fluctuations of pain within subjects and physiological markers of pain: Donaldson et al. 2003). Corresponding remarks apply to the case of pleasure (Haybron 2008).

Philosophers arguing against infallibilism have concocted hypothetical examples in which they argue it is plausible to attribute introspective error; but even if the examples

succeed, they are generally confined to far-fetched scenarios, pathological cases, or very minor or very brief mistakes (Armstrong 1963; Churchland 1988; Kornblith 1998, with an eye to the distinction between mistakes about current conscious experience and other sorts of mistakes). A different kind of disconnection between conscious experience and introspective judgment is suggested by Block (2007), who argues that phenomenal consciousness contains much more detail than we can report. For example, according to Block, when we take in a complex visual scene we consciously experience a wealth of detail only a small portion of which can go into working memory for the sake of introspective report.

In contrast with the dominant philosophical tradition that stresses the special privilege or at least high accuracy of introspective judgments about consciousness, many early introspective psychologists held that the introspection of currently ongoing or recently past conscious experience is difficult and prone to error if the introspective observer is insufficiently trained. Wundt, for example, who is generally regarded as the founder of experimental psychology, reportedly did not credit the introspective reports of individuals with fewer than 50,000 trials of practice in observing their conscious experience (Boring 1953). Titchener, perhaps the leading American introspective psychologist, designed a 1600-page training manual for students, arguing that introspective observation is at least as difficult as observation in the physical sciences (Titchener 1901-1905; see also Müller 1904). However, such early introspective training did not succeed in generating dependably consistent reports across laboratories, for instance on issues such as the existence of "imageless thought" (Humphrey 1951). Such persistent disagreements

contributed to the notorious demise of introspective psychology in favor of behaviorist methods which, by confining themselves to the measure of easily observable stimuli and behavior, yielded results that replicated more dependably across laboratories (Samelson 1981; Mills 1998; but see Hurlburt 1993; for more recent discussions of introspective training see Varela 1996; Nahmias 2002; Schwitzgebel 2004).

Ericsson and Simon (1984/1993) discuss and review relationships between performance on various problem-solving tasks, concurrent verbalizations of thoughts ("think aloud protocols"), and immediately retrospective verbalizations. The existence of good relationships in the predicted directions in many problem-solving tasks lends empirical support to the view that people's reports about their stream of thoughts often accurately reflect those thoughts. (In the case of concurrent verbalization, such accuracy fits very naturally with self-fulfillment or expressivist models of self-knowledge; see Sections 2.3.2 [*insert link] and 2.3.3 [*insert link].) For example, Ericsson and Simon find that think-aloud and retrospective reports of thought processes correlate with predicted patterns of eye movement and response latency (see also Ericsson 2003). Ericsson and Simon also cite studies like that of Hamilton and Sanford (1978), who asked subjects to make yes or no judgments about whether pairs of letters were in alphabetical order (like MO) or not (like RP) and then to describe retrospectively their method for arriving at the judgments. When subjects retrospectively reported knowing the answer "automatically" without an intervening conscious process, reaction times were swift and did not depend on the distance between the letters. When subjects retrospectively reported "running through" a sequential series of letters (such as "LMNO" when prompted with "MO")

reaction times correlated nicely with reported length of run-through. On the other hand, Flavell, Green, and Flavell (1995) report gross and systematic introspective error about recently past and even current (conscious) thought in young children; and Smallwood and Schooler (2006) review literature that suggests that people are not especially good at detecting when their mind is wandering.

The most sustained case against the accuracy of introspective reports about consciousness has been made by Schwitzgebel. Some of Schwitzgebel's arguments turn on arguing that conscious experience has a particular property that most people would deny it has. Schwitzgebel argues, for example, that introspection tends to mislead people about the level of detail and clarity in the visual field (2008; see also Dennett 1991; Blackmore 2002). He asserts that most people when queried will report a visual experience of stable clarity over thirty degrees or more of the visual field, while physiological and behavioral evidence, as well as the evidence of more careful introspection, suggests that instead visual experience involves a very small region of clarity (1-2 degrees of visual arc) moving rapidly around a hazy background. Schwitzgebel also claims, again on the basis of behavioral and introspective evidence, that ordinary people auditorially experience the echoic properties of silent objects, though they typically deny having such experiences (Schwitzgebel and Gordon 2000), and that due to overanalogizing to flat media like paintings and photographs, contemporary Westerners mistakenly attribute "projective distortions" to visual experience, such as in claiming that an obliquely-viewed coin looks elliptical (Schwitzgebel 2006).

Other of Schwitzgebel's arguments turn on cases of radical intersubjective disagreement about conscious experience where due to physiological and behavioral similarity he claims it is plausible to assume that people's experiences are similar. For example, he points to philosophical disagreements about whether there is a "phenomenology of thinking" beyond that of imagery and emotion (Schwitzgebel 2008); disagreements about whether sensory experience is "rich" (including for example constant tactile experience of one's feet in one's shoes) or "thin" (limited mostly just to what is in attention at any one time) (Schwitzgebel 2007a); and disagreements about the degree of vividness in visual imagery (which Schwitzgebel argues is poorly correlated with behavior on tasks psychologists have traditionally thought to involve imagery; Schwitzgebel 2002a). However, Schwitzgebel's argument on the last point is at least partly undermined by recent neurophysiological studies that find differences in subjects' cortical activation that appear to vary in systematic, predictable ways along with their introspective reports (Amedi, Malach, and Pascual-Leone 2005; Cui et al. 2007).

3.4. Self-Knowledge and Content Externalism

Finally, the large literature on the relationship between self-knowledge and content externalism merits mention. Content externalism is the view that the particular content of a person's attitude (such as the belief that alcohol boils at 78 degrees Celsius) depends not just causally but constitutively on how things stand in the external environment (such as the alcohol around her), such that is it possible that two people who are molecule-for-molecule identical with each other may nonetheless have different attitudes (e.g., Putnam

1975; Burge 1979). Adapting and example from Putnam, imagine a "Twin Earth" in a far-away galaxy which is molecule-for-molecule identical with Earth, except that where we have alcohol they have *twalcohol*, a substance with all the same superficial characteristics and effects of alcohol but a different chemical formula. Jacqui on Earth and her twin on Twin Earth may both say and think the form of words "alcohol boils at 78 degrees"; but the content of Jacqui's thought may be that alcohol does, while the content of her twin's thought is that twalcohol does. Boghossian (1989) and others have argued that if content externalism is true, then there should be thoughts with different contents that are introspectively identical, and thus (implausibly, Boghossian thinks) people should not be able to know the content of their attitudes by introspection alone. Containment [*insert link] and transparency [*insert link] views of introspection are sometimes developed partly in reaction to arguments of this sort (e.g., Burge 1988; Heil 1988; Dretske 1995). The topic of content externalism and self-knowledge is explored in further in the entries on self-knowledge and externalism about mental content.

Bibliography

- Alston, William P. (1971/1989). Varieties of privileged access, *American Philosophical Quarterly, 8,* 223-241. Reprinted in William P. Alston, *Epistemic justification.* Ithaca, NY: Cornell.
- Amedi, Amir, Rafael Malach, and Alvaro Pascual-Leone (2005). Negative BOLD differentiates visual imagery and perception. *Neuron, 48,* 859-872.
- Armstrong, David M. (1963). Is introspective knowledge incorrigible? *Philosophical Review, 72,* 417-432.

- ----- (1968). *A materialist theory of the mind.* London: Routledge.
- ----- (1980). *The nature of mind and other essays.* Ithaca, NY: Cornell.
- ----- (1999). *The mind-body problem.* Boulder, CO: Westview.
- Aronson, Elliott (1968). Dissonance theory: Progress and problems. In Robert P. Ableson et al. (eds.), *Theories of cognitive consistency*. Chicago: Rand McNally.
- Augustine, Aurelis (c. 420 C.E./1998). *The city of god against the pagans*, trans. R.W. Dyson. Cambridge: Cambridge.
- Aydede, Murat, and Güven Güzeldere (2005). Cognitive architecture, concepts, and introspection: An information-theoretic solution to the problem of phenomenal consciousness. *Nous, 39,* 197-255.
- Ayer, A.J. (1936/1946). *Language, truth, and logic, 2nd ed.* London: Gollancz.
- ----- (1956). *The problem of knowledge.* London: Macmillan.
- Baldwin, James Mark (1901-1905). *Dictionary of philosophy and psychology.* New York: Macmillan.
- Bem, Daryl J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74,* 183-200.
- ----- (1972). Self-perception theory. *Advances in Experimental Social Psychology, 6,* 1-62.
- Blackmore, Susan (2002). There is no stream of consciousness. *Journal of Consciousness Studies, 9* (no. 5-6), 17-28.
- Block, Ned (1995/1997). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18,* 227-247. Revised and reprinted in Ned Block,

Güven Güzeldere, and Owen Flanagan, eds., *The nature of consciousness.* Cambridge, MA: MIT.

- ----- (1996). Mental paint and mental latex. *Philosophical Issues, 7,* 19-49.
- ----- (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences, 30,* 481-499.
- Boghossian, Paul (1989). Content and self-knowledge. *Philosophical Topics, 17,* 5-26.
- Boring, E.G. (1953). A history of introspection. *Psychological Bulletin, 50,* 169-189.
- Brentano, Franz (1874/1995). *Psychology from an empirical standpoint.* Trans. Antos C. Rancurello, D. B. Terrell and Linda L. McAlister, 2nd English edition. New York: Routledge.
- Burge, Tyler (1979). Individualism and the mental. In *Midwest Studies in Philosophy, 4,* 73-121.
- ----- (1988). Individualism and self-knowledge. *Journal of Philosophy, 85,* 649-663.
- ----- (1996). Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society, 96,* 91-116.
- ---- (1998). Reason and the first person. In Crispin Wright, Barry C. Smith, and Cynthia Macdonald (eds.), *Knowing our own minds.* Oxford: Oxford.
- Byrne, Alex (1995). Introspection. *Philosophical Topics, 33* (no. 1), 79-104.
- Campbell, Donald T., William H. Kruskal, and William P. Wallace (1966). Seating aggregation as an index of attitude. *Sociometry, 29,* 1-15.
- Campbell, John (1999). Immunity to error through misidentification and the meaning of a referring term. *Philosophical Topics, 25* (no. 1-2), 89-104.

- Carruthers, Peter (2005). *Consciousness: Essays from a higher-order perspective.* Oxford: Oxford.
- ----- (forthcoming-a). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences.*
- ----- (forthcoming-b). Introspection: Divided and partly eliminated. *Philosophy and Phenomenological Research.*
- Cavell, Marcia (2006). *Becoming a subject.* Oxford: Oxford.
- Chalmers, David J. (1996). *The conscious mind.* New York: Oxford.
- ----- (2003). The content and epistemology of phenomenal belief. In Quentin Smith and Aleksandar Jokic (eds.), *Consciousness: New philosophical perspectives.* Oxford: Oxford.
- Chisholm, Roderick M. (1981). *The first person.* Brighton, UK: Harvester.
- Churchland, Paul M. (1988). *Matter and consciousness, rev. ed.* Cambridge, MA: MIT.
- Comte, Auguste (1830). *Cours de philosophie positive, vol. 1.* Paris: Bacheleier, Libraire pour les Mathématiques.
- Cooper, Joel, and Russell H. Fazio (1984). A new look at dissonance theory. *Advances in Experimental Social Psychology, 17,* 229-266.
- Cui, Xu, Cameron B. Jeter, Dongni Yang, P. Read Montague, and David M. Eagleman (2007). Vividness of mental imagery: Individual variability can be measured objectively. *Vision Research, 47,* 474-478.
- Dennett, Daniel C. (1987). *The intentional stance.* Cambridge, MA: MIT.
- ----- (1991). *Consciousness explained.* Boston: Little, Brown, and Co.

- ----- (2000). The case for rorts. In R. B. Brandom (ed.), *Rorty and his critics*. Malden, MA: Blackwell.
- ----- (2002). How could I be wrong? How wrong could I be? *Journal of Consciousness Studies, 9* (no. 5-6), 13-6.
- Descartes, René (1637/1985). *Discourse on the method*. In *The philosophical writings of Descartes, vol. 1,* ed. and trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge.
- ----- (1641/1984). *Meditations on first philosophy*. In *The philosophical writings of Descartes, vol. 2,* ed. and trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge.
- Donaldson, Gary W., C. Richard Chapman, Yoshi Nakamure, David H. Bradshaw, Robert C. Jacobson, and Christopher N. Chapman (2003). Pain and the defense response: Structural equation modeling reveals a coordinated psychophysiological response to increasing painful stimulation. *Pain, 102,* 97-108.
- Dovidio, John F., Kerry Kawakami, Craig Johnson, Brenda Johnson, and Adaiah Howard (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology, 33,* 510-540.
- Dretske, Fred (1995). *Naturalizing the mind.* Cambridge, MA: MIT.
- ---- (2004). Knowing what you think vs. knowing that you think it. In Richard Schantz (ed.), *The externalist challenge.* Berlin: Walter de Gruyter.
- Ericsson, K. Anders, and Herbert A. Simon (1984/1993). *Protocol analysis, rev. ed.* Cambridge, MA: MIT.

- Ericsson, K. Anders (2003). Valid and non-reactive verbalization of thoughts during performance of tasks: Towards a solution to the central problems of introspection as a source of scientific data. *Journal of Consciousness Studies, 10* (no. 9-10), 1-18.
- Evans, Gareth (1982). *Varieties of reference.* Ed. John McDowell. Oxford: Oxford.
- Fazio, Russell H., Joni R. Jackson, Bridget C. Dunton, and Carol J. Williams (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013-1027.
- Fernández, Jorgi (2003). Privileged access naturalized. *Philosophical Quarterly, 53,* 352-372.
- Festinger, Leon (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford.
- Festinger, Leon, and James M. Carlsmith (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58,* 203-210.
- Flavell, John H., Frances L. Green, and Eleanor R. Flavell (1995). *Young children's knowledge about thinking. Monographs of the Society for Research in Child Development, 60* (no. 243).
- Fodor, Jerry A. (1983). *Modularity of mind.* Cambridge, MA: MIT.
- ----- (1998). *Concepts: Where cognitive science went wrong.* Oxford: Oxford.
- Fumerton, Richard (2006). Direct realism, introspection, and cognitive science. *Philosophy and Phenomenological Research, 73,* 680-695.
- Funder, David C. (1999). *Personality judgment.* London: Academic.
- Gazzaniga, Michael S. (1995). Consciousness and the cerebral hemispheres. In Michael S. Gazzaniga (ed.), *The Cognitive Neurosciences.* Cambridge, MA: MIT.

- Gawronski, Bertram, and Galen V. Bodenhausen (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731.
- Gendler, Tamar Szabó (forthcoming-a). Alief and Belief. *Journal of Philosophy.*
- ----- (forthcoming-b). Alief in Action (and Reaction). *Mind & Language.*
- Gennaro, Rocco J. (1996). *Consciousness and Self-Consciousness*. Amsterdam: John Benjamins.
- Gertler, Brie (2000). The mechanics of self-knowledge. *Philosophical Topics, 28,* 125-146.
- ----- (2001). Introspecting phenomenal states. *Philosophy and Phenomenological Research, 63,* 305-328.
- ----- (forthcoming). Self-knowledge and the transparency of belief. In Anthony Hatzimoysis (ed.), *Self-knowledge.* Oxford: Oxford.
- Goldman, Alvin I. (1989). Interpretation psychologized. *Mind and Language, 4,* 161-185.
- ----- (2000). Can science know when you're conscious? *Journal of Consciousness Studies, 7* (no. 5), 3-22.
- ----- (2006). *Simulating minds.* Oxford: Oxford.
- Gopnik, Alison (1993a). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16,* 1-14.
- ----- (1993b). Psychopsychology. *Consciousness and Cognition, 2,* 264-280.
- Gopnik, Alison, and Andrew N. Meltzoff (1994). Minds, bodies and persons: Young children's understanding of the self and others as reflected in imitation and "theory of

mind" research. In Sue Taylor Parker, Robert W. Mitchell, and Maria L. Boccia (eds.), *Self-awareness in animals and humans.* New York: Cambridge.

- Gordon, Robert M. (1986). Folk psychology as simulation. *Mind and Language, 1,* 151-171.
- ----- (1995). Simulation without introspection or inference from me to you. In Martin Davies and Tony Stone, eds., *Mental simulation.* Oxford: Blackwell.
- ----- (2007). Ascent routines for propositional attitudes. *Synthese, 159,* 151-165.
- Greenwald, Anthony G., Debbie E. McGhee, and Jordan L.K. Schwartz (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464-1480.
- Greenwald, Anthony G., and Brian A. Nosek (2009). Attitudinal dissociation: What does it mean? In Richard E. Petty, Russell H. Fazio, and Pablo Briñol (eds.), *Attitudes: Insights from the New Implicit Measures.* New York: Taylor and Francis.
- Hamilton, Andy (2008). Memory and self-consciousness: Immunity to error through misidentification. *Synthese,* online first: DOI 10.1007/s11229-008-9318-6
- Hamilton, J.M.E., and A.J. Sanford (1978). The symbolic distance effect for alphabetic order judgements: A subjective report and reaction time analysis. *Quarterly Journal of Experimental Psychology, 30,* 33-43.
- Harman, Gilbert (1990). The intrinsic quality of experience. In James Tomberlin (ed.), *Philosophical Perspectives, 4.* Ridgeview.
- Haybron, Daniel M. (2008). *The pursuit of unhappiness.* Oxford: Oxford.
- Hecker, Joel M., Jennifer A. Schmidt, and Mihaly Csikszentmihalyi (2007). *Experience sampling method.* Thousand Oaks, CA: Sage.

- Heil, John (1988). Privileged access. *Mind, 97,* 238-251.
- Hill, Christoper S. (1991). *Sensations: A defense of type materialism.* Cambridge: Cambridge.
- ----- (forthcoming). *Consciousness.* Cambridge: Cambridge.
- Hirstein, William (2005). *Brain fiction.* Cambridge, MA: MIT.
- Horgan, Terence, and John Tienson (2002). The intentionality of phenomenology and the phenomenology of intentional. In David J. Chalmers (ed.), *Philosophy of Mind.* Oxford: Oxford.
- Horgan, Terence, John Tienson, and George Graham (2006). Internal-world skepticism and the self-presentational nature of phenomenal consciousness. In Uriah Krigel and Kenneth Williford, eds., *Self-representational approaches to consciousness.* Cambridge, MA: MIT.
- Horgan, Terence, and Uriah Kriegel (2007). Phenomenal epistemology: What is consciousness that we may know it so well? *Philosophical Issues, 17,* 123-144.
- Hume, David (1739/1978). *A treatise of human nature,* edited by L.A. Selby-Bigge and P.H. Nidditch. Oxford: Clarendon.
- ----- (1748/1975). *An enquiry concerning human understanding.* In *enquiries concerning human understanding and concerning the principles of morals*, edited by L.A. Selby-Bigge and P.H. Nidditch. Oxford: Clarendon.
- Humphrey, George (1951). *Thinking: An introduction to its experimental psychology.* London: Methuen.
- Hurlburt, Russell T. (1990). *Sampling normal and schizophrenic inner experience.* New York: Plenum.

- Hurlburt, Russell T. and Christopher L. Heavey (2006). *Exploring inner experience.* Amsterdam: John Benjamins.
- Hurlburt, Russell T. and Eric Schwitzgebel (2007). *Describing inner experience? Proponent meets skeptic*. Cambridge, MA: MIT.
- Ito, Tiffany A., and John T. Cacioppo (2007). Attitudes as mental and neural states of readiness. In Bernd Wittenbrink and Norbert Schwarz (eds.), *Implicit measures of attitudes.* New York: Guilford.
- Jackson, Frank (1977). *Perception.* Cambridge: Cambridge.
- James, William (1890/1981). *The principles of psychology.* Cambridge, MA: Harvard.
- Jaynes, Julian (1976). *The origin of consciousness in the breakdown of the bicameral mind.* New York: Houghton Mifflin.
- Johansson, Petter, Lars Hall, Sverker Sikström, and Andreas Olsson (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310,* 116-119.
- Johansson, Petter, Lars Hall, Sverker Sikström, Betty Tärning, and Andreas Lind (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition, 15,* 673-692.
- Kant, Immanuel (1781/1997). *The critique of pure reason.* Ed. and trans. Paul Guyer and Allen W. Wood. Cambridge: Cambridge.
- Kay, Aaron C., Maria C. Jimenez, and John T. Jost (2002). Sour grapes, sweet lemons, and the anticipatory rationalization of the status quo. *Personality and Social Psychology Bulletin, 28,* 1300-1312.

- Kihlstrom, John F. Implicit methods in social psychology. In Carol Sansone, Carolyn C. Morf, and A.T. Panter, eds., *The SAGE handbook of methods in social psychology.* Thousand Oaks, CA: Sage.
- Kind, Amy (2003). What's so transparent about transparency? *Philosophical Studies, 115,* 225-244.
- Kornblith, Hilary (1998). What is it like to be me? *Australasian Journal of Philosophy, 76,* 48-60.
- Kriegel, Uriah (2006). The same-order monitoring theory of consciousness. In Uriah Krigel and Kenneth Williford, eds., *Self-representational approaches to consciousness.* Cambridge, MA: MIT.
- Lane, Kristin A., Mahzarin R. Banaji, Brian A. Nosek, and Anthony G. Greenwald (2007). Understanding and using the Implicit Association Test: IV. In Bernd Wittenbrink and Norbert Schwarz (eds.), *Implicit measures of attitudes.* New York: Guilford.
- Larson, Reed, and Mihaly Csikszentmihalyi (1983). The Experience Sampling Method. In Harry T. Reis (ed.), *Naturalistic approaches to studying social interaction*. San Francisco: Jossey-Bass.
- Lear, Jonathan (1998). *Open-minded.* Cambridge, MA: Harvard.
- Lewis, C.I. (1946). *An analysis of knowledge and valuation.* La Salle, IL: Open Court.
- Locke, John (1690/1975). *An essay concerning human understanding.* Ed. Peter H. Nidditch. Oxford: Oxford.
- Lycan, William G. (1996). *Consciousness and experience.* Cambridge, MA: MIT.

- Lyons, William (1986). *The disappearance of introspection.* Cambridge, MA: MIT.
- Lyubomirsky, Sonja, and Lee Ross (1999). Changes in attractiveness of elected, rejected, and precluded alternatives: A comparison of happy and unhappy individuals. *Journal of Personality and Social Psychology, 76,* 988-1007.
- Marr, David (1983). *Vision.* New York: Freeman.
- Martin, Michael G.F. (2002). The transparency of experience. *Mind and Language, 17,* 376-425.
- Maudsley, Henry (1867/1977). *Physiology and pathology of the mind.* Ed. Daniel N. Robinson. Washington, DC: University Publications of America.
- McGeer, Victoria (1996). Is "self-knowledge" an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy, 93,* 483-515.
- McGeer, Victoria, and Philip Pettit (2002). The self-regulating mind. *Language and Communication, 22,* 281-299.
- Mele, Alfred (2001). *Self-deception unmasked.* Princeton, NJ: Princeton.
- Mill, John Stuart (1865/1961). *Auguste Comte and positivism.* Ann Arbor, MI: University of Michigan.
- Mills, John A. (1998). *Control: A history of behavioral psychology.* New York: NYU.
- Moll, Albert (1889/1911). *Hypnotism.* Ed. Arthur F. Hopkirk. New York: Charles Scribner's Sons.
- Moore, George Edward (1903/1922). The refutation of idealism. *Mind, 12,* 433-453. Reprinted in George Edward Moore, *Philosophical studies.* London: George Allen & Unwin.

- ----- (1942). A reply to my critics. In P. A. Schlipp (ed.), *The philosophy of G. E. Moore*. New York: Tudor.
- ----- (1944/1993). Moore's paradox. In G.E. Moore, *Selected writings,* ed. Thomas Baldwin. London: Routledge.
- Moran, Richard (2001). *Authority and estrangement.* Princeton: Princeton.
- Müller, G.E. (1904). *Die Gesichtspunkte und die Tatsachen der psychophysischen Methodik.* Wiesbaden: J.F. Bergmann.
- Nahmias, Eddy (2002). Verbal reports on the contents of consciousness: Reconsidering introspectionist methodology. *Psyche, 8* (no. 21).
- Nisbett, Richard E., and Nancy Bellows (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of Personality and Social Psychology, 35,* 613-624.
- Nisbett, Richard E., and Lee Ross (1980). *Human inference.* Englewood Cliffs, NJ: Prentice-Hall.
- Nisbett, Richard E., and Timothy DeCamp Wilson (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.
- Noë, Alva (2004). *Action in perception.* Cambridge, MA: MIT.
- Papineau, David (2002). *Thinking about consciousness.* Oxford: Oxford.
- Peacocke, Christopher (1998). Conscious attitudes, attention, and self-knowledge. In Crispin Wright, Barry C. Smith, and Cynthia Macdonald (eds.), *Knowing our own minds.* Oxford: Oxford.
- ----- (1999). *Being known.* Oxford: Oxford.

- Petty, Richard E., Russell H. Fazio, and Pablo Briñol, eds. (2009). *Attitudes: Insights from the New Implicit Measures.* New York: Taylor and Francis.
- Pitt, David (2004). The phenomenology of cognition, or what is it like to think that P? *Philosophy and Phenomenological Research, 69,* 1-36.
- Prinz, Jesse (2004). The fractionation of introspection. *Journal of Consciousness Studies, 11* (no. 7-8), 40-57.
- Pryor, James (1999). Immunity to error through misidentification. *Philosophical Topics, 26* (no. 1-2), 271-304.
- Putnam, Hilary (1975). The meaning of 'meaning'. In Hilary Putnam, *Philosophical papers, vol. 2.* Cambridge: Cambridge.
- Redelmeier, Donald A., and Daniel Kahneman (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66,* 3-8.
- Richet, Charles (1884). *L'homme et l'intelligence.* Paris: F. Alcan.
- Robinson, William S. (2005). Thoughts without distinctive non-imagistic phenomenology. *Philosophy and Phenomenological Research, 70*, 534-60.
- Rorty, Richard (1970). Incorrigibility as the mark of the mental. *Journal of Philosophy, 67*, 399–424.
- Rosenthal, David M. (1990). Two concepts of consciousness. *Philosophical Studies*, *49,* 329-359
- ----- (2001/2005). Introspection and self-interpretation. *Philosophical Topics, 28* (no. 2), 201-233. Reprinted in David M. Rosenthal, *Consciousness and Mind.* Oxford: Oxford.

- ----- (2005). *Consciousness and Mind.* Oxford: Oxford.
- Ryle, Gilbert (1949). *The concept of mind.* New York: Barnes and Noble.
- Samelson, Franz (1981). Struggle for scientific authority: The reception of Watson's behaviorism, 1913-1920. *Journal of the History of the Behavioral Sciences, 17,* 399-425.
- Schwitzgebel, Eric (2002a). A phenomenal, dispositional account of belief. *Nous, 36,* 249-275.
- Schwitzgebel, Eric (2002b). How well do we know our own conscious experience? The case of visual imagery. *Journal of Consciousness Studies, 9* (no. 5-6), 35-53.
- ----- (2004). Introspective training apprehensively defended: Reflections on Titchener's lab manual. *Journal of Consciousness Studies, 11* (no. 7-8), 58-76.
- ----- (2006). Do things look flat? *Philosophy & Phenomenological Research, 72,* 589-599.
- ----- (2007a). Do you have constant tactile experience of your feet in your shoes? Or is experience limited to what's in attention? *Journal of Consciousness Studies, 14* (no. 3), 5-35.
- ----- (2007b). No unchallengeable epistemic authority, of any sort, regarding our own conscious experience -- contra Dennett? *Phenomenology and the Cognitive Sciences, 6,* 107-113.
- ----- (2008). The unreliability of naive introspection. *Philosophical Review, 117,* 245-273.

- Schwitzgebel, Eric and Michael S. Gordon (2000). How well do we know our own conscious experience? The case of human echolocation. *Philosophical Topics, 28,* 235-246.
- Searle, John R. (1992). *The rediscovery of the mind.* Cambridge, MA: MIT.
- Shoemaker, Sydney (1963). *Self-knowledge and self-identity.* Ithaca, NY: Cornell.
- ----- (1968/2003). Self-reference and self-awareness. *Journal of Philosophy, 65,* 555-567. Reprinted in Sydney Shoemaker, *Identity, Cause, and Mind, expanded ed.* Oxford: Oxford.
- ----- (1988/1996). On knowing one's own mind. *Philosophical Perspectives, 2,* 183-209. Reprinted in Sydney Shoemaker, *The first-person perspective and other essays.* Cambridge: Cambridge.
- ----- (1994a/1996). Self-knowledge and "inner sense." Lecture I: The object perception model. Reprinted in Sydney Shoemaker, *The first-person perspective and other essays.* Cambridge: Cambridge.
- ----- (1994b/1996). Self-knowledge and "inner sense." Lecture II: The broad perceptual model. Reprinted in Sydney Shoemaker, *The first-person perspective and other essays.* Cambridge: Cambridge.
- ----- (1994c/1996). Self-knowledge and "inner sense." Lecture III: The phenomenal character of experience. Reprinted in Sydney Shoemaker, *The first-person perspective and other essays.* Cambridge: Cambridge.
- Siewert, Charles (1998). *The significance of consciousness.* Princeton, NJ: Princeton.
- ----- (2004). Is experience transparent? *Philosophical Studies, 117,* 15-41.

- Sirken, Monroe G., Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith N. Tanur, Roger Tourangeau, eds., (1999). *Cognition and survey research.* New York: John Wiley and Sons.
- Smallwood, Jonathan, and Jonathan W. Schooler (2006). The restless mind. *Psychological Bulletin, 132,* 946-958.
- Smith, A.D. (2008). Translucent experiences. *Philosophical Studies, 140,* 197-212.
- Stoljar, Daniel (2004). The argument from diaphanousness. In Maite Ezcurdia, Robert J. Stainton, and Christopher Viger (eds.), *New essays in the philosophy of language and mind.* Calgary: University of Calgary.
- Stone, Jeff, and Joel Cooper (2001). A self-standards model of cognitive dissonance. *Journal of Experimental Social Psychology, 37,* 228-243.
- Taylor, Shelley E., and Jonathon D. Brown (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193-210.
- Thomas, Nigel (1999). Are theories of imagery theories of imagination? *Cognitive Science, 23,* 207-45.
- Titchener, E.B. (1901-1905). *Experimental psychology.* New York: Macmillan.
- ----- (1912a). Prolegomena to a study of introspection. *American Journal of Psychology, 23,* 427-448.
- ----- (1912b). The schema of introspection. *The American Journal of Psychology, 23,* 485-508.
- Tye, Michael (1995). *Ten problems about consciousness.* Cambridge, MA: MIT.
- ----- (2000). *Consciousness, color, and content.* Cambridge, MA: MIT.

- ----- (2002). Representationalism and the transparency of experience. *Nous, 36,* 137-151.
- Van Gulick, Robert (1993). Understanding the phenomenal mind: Are we all just armadillos? In Martin Davies and Glyn W. Humphreys (eds.), *Consciousness: Psychological and philosophical essays*. Oxford: Blackwell.
- Vanman, Eric J., Brenda Y. Paul, Tiffany A. Ito, and Norman Miller (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology, 73,* 941-959.
- Varela, Francisco J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies, 3* (no. 4)*,* 330-49.
- Velleman, J. David (2000). *The possibility of practical reason.* Oxford: Oxford.
- Wegner, Daniel M. (2002). *The illusion of conscious will.* Cambridge, MA: MIT.
- Wegner, Daniel M. and Thalia Wheatley (1999). Apparent mental causation. *American Psychologist, 54,* 480-492.
- White, Peter A. (1988). Knowing more about what we can tell: "Introspective access" and causal report accuracy ten years later. *British Journal of Psychology, 79,* 13-45.
- Williams, Amanda C. de C., Huw Talfryn Oakley Davies, and Yasmin Chadury (2000). Simple pain rating scales hide complex idiosyncratic meanings. *Pain, 85,* 457-463.
- Wilson, Timothy D. (2002). *Strangers to ourselves.* Cambridge, MA: Harvard.
- Wilson, Timothy D., Samuel Lindsey, and Tonya T. Schooler (2000). A model of dual attitudes. *Psychological Review, 107,* 101-126.

- Wittenbrink, Bernd, Charles M. Judd, and Bernadette Park (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72,* 262-274.
- Wittenbrink, Bernd, and Norbert Schwarz, eds. (2007). *Implicit measures of attitudes.* New York: Guilford.
- Wittgenstein, Ludwig (1953/1968). *Philosophical investigations, 3rd ed.* Trans. G.E.M. Anscombe. New York: Macmillan.
- Wollheim, Richard (1981). *Sigmund Freud.* New York: Cambridge.
- ----- (2003). On the Freudian unconscious. *Proceedings and Addresses of the American Philosophical Association, 77* (no. 2), 23-35.
- Wright, Crispin (1989). Wittgenstein's later philosophy of mind: Sensation, privacy, and intention. *Journal of Philosophy, 86,* 622-634.
- ----- (1998). Self-knowledge: The Wittgensteinian legacy. In Crispin Wright, Barry C. Smith, and Cynthia Macdonald (eds.), *Knowing our own minds.* Oxford: Oxford.
- Zimmerman, Aaron (2007). The nature of belief. *Journal of Consciousness Studies,* 14 (no. 11), 61-82.