Introspection, What?

Eric Schwitzgebel
Department of Philosophy
University of California at Riverside
Riverside, CA 92521-0201

eschwitz at domain- ucr.edu

March 30, 2010

Introspection, What?

1. Introduction.

In this essay I argue for two theses. One is that introspection is a species of attention to conscious experience, one that aims to exhibit what I call *relatively direct sensitivity* to the experience. The other is that introspection is not the operation of a single, dedicated mechanism or family of dedicated mechanisms (such as self-scanning or self-monitoring devices); rather, in introspecting we opportunistically deploy a variety of cognitive systems and processes.

The following analogy might be helpful. Suppose you are at a psychology conference or a high school science fair and you are trying to quickly take in a poster. You are not equipped with a dedicated faculty of *poster-taking-in*. Rather, you opportunistically deploy a variety of processes with the aim of getting the gist of the poster: You look at the poster; you listen to what the poster's author is saying; you follow out implications, charitably rejecting some interpretations of the poster's content as too obviously foolish; you think about what it makes sense to claim given the social and scientific context and other work by the author or the author's advisor, if you know any; you pose questions and assess the author's responses both for overt content and for emotional flavor. Although the cognitive systems involved range widely and are not dedicated just to taking in posters, not just any activity counts as taking in a poster: You must devote attention to the task, and you must do so in a way that involves a certain sort of visual responsiveness to the contents of the poster. If visual information from the poster has no effect on your judgments about the poster's contents, then you may be listening to the author, but you are not taking in the poster. Similarly, introspection requires attention and involves a wide variety of cognitive systems that are not inherently introspective, and yet not just any attention-

consuming process that results in a judgment about your own mind is an introspective process. Taking in the poster requires visual sensitivity to it; analogously, successful introspection requires, as I will explain, relatively direct sensitivity to the target experience.

2. Examples of Introspection.

My account of introspection emphasizes the complexity and diversity of the cognitive processes generating introspective judgments. Most philosophers, when they discuss the nature and epistemology of introspection, use only maximally simple examples ("I'm in pain", "I have a visual experience of red", "I believe P"). Sometimes, perhaps, introspective judgments are very simple – but most introspective activity, I suspect, is rather more complex, both in everyday life and for the purposes of psychology and consciousness studies. It is methodologically dubious, then, to build a general account of introspection from consideration of only a few very simple or highly abstract examples. Consider, then, these somewhat more complex examples:

(1.) *Looking at a tree.* I gaze out my window at a tree. I don't simply notice the tree and its features; I also think about my visual experience of the tree. I notice that I am seeing the greenness of the leaves and the way the water from a recent rain seems to sparkle in the sun. I notice that when I fix my gaze on the leftmost lower branch of the tree, the top of the tree is not visually clear. I notice that the green coloration of the leaves in the shadows looks in some ways similar to and in other ways different from the green coloration of the leaves in direct sunlight.

(2.) *Assessing my current emotional experience.* I think about what, if anything, I am emotionally experiencing right now. I notice, first, that my lips are pursed, and I relax them; I notice some tension in my chest. But then I think to myself that emotional experience is not, or might not be, all about the body. Some kind of negative affect is present – perhaps I'm tense

about writing this essay? But I had been looking forward all day to finally having the time to do it! I am tense about the time, I decide, and the looming deadline. I find my lips pursed again and rub them with my left hand. Being tense about the deadline doesn't seem like the only thing that is going on with me emotionally right now – but what more there is I can't quite put a finger on. My throat is tense and my brows are furrowed. I think that thinking about this particular introspective task is worsening my mood, making me tenser and maybe almost angry. I would have liked a happier example, one that better displays the sunny disposition and amiable character I believe myself to have. Perhaps, partly, I am distressed at the negativity of this example, and that distress is further reinforcing the negativity.

(3.) *Noticing a tune in my head.* In the morning on the way to work, I had been blasting a tune on the stereo, Sonic Youth's "Kool Thing". Now it's two in the afternoon and I notice that that tune has been running through my head. My sense is that it has been running for at least three seconds, maybe much longer, and not for the first time today. As I reflect, the tune seems to sharpen up or become more vivid. It seems that I can choose to emphasize the vocals or the guitars, and I think about the extent to which I can imagine both the vocals and guitars simultaneously. It seems to me that I can do so, especially if I nod my head in rhythm and do something that feels like using my mouth and voice to track the lead guitar line (though no noticeable sound issues from my mouth).

(4.) *Deciding about the pie.* Apple pie and cherry pie are on the menu, and I must decide which I prefer. I try to call up in memory the taste of each. I think about the fact that although I greatly prefer cherries to apples, cherry pie has regularly disappointed me in the past. I imagine the crunchy crumble layer that tops Dutch apple pie, but then scold myself because Dutch apple isn't what's on the menu. Nonetheless, I figure that since the restaurant isn't first-rate I'd prefer

not to risk the cherry – and as I make that decision, I know, or at least believe, that the primary basis of my preference is my guess that cherry would probably only disappoint.

I hope you'll agree that these are examples of introspection. If you're willing to take examples like this as paradigmatic starting points, and if you think that in making these examples complex I haven't buried the introspective process amidst a pile of non-introspective processes but rather illustrated the complexity often inherent in introspective activity – well, then you have already come a fair way down the road I wish to lead you along.

In my jargon, the *target* of the introspective process is the mental state or mental process or mental event (from now on, I will just say either "state" or "process" as seems most natural, without, I hope, losing sight of the fact that these are important differences in conception) about which one aims to reach an introspective judgment. In three of the four examples above, the target is a current (or maybe very recently past) conscious experience. I have drawn one example from each of the three major types of conscious experience that are nearly universally acknowledged: sensory experience (in this case, visual experience of a tree), imagery experience (in this case, auditory or mostly auditory imagery of a tune), and emotional experience. In the fourth case – deciding about the pie – how exactly to conceive of the introspective target or targets is rather more vexed, but I hope you'll agree that introspection is involved in some way. All four examples will be developed further below.

3. Self-Knowledge Without Introspective Self-Detection.

Existing general accounts of introspection, and of self-knowledge of one's own current (or very recently past) mental states, tend to fall into one of two types. The truth, I want to suggest, lies in the middle, but no theorist adequately occupies the middle. One type of account

treats introspection as the operation of a self-detection process. The other type of account focuses on methods for generating self-knowledge without introspective self-detection.

Introspective self-detection accounts treat the target state and the introspective judgment (or introspective belief, or whatever else is supposed to be the product of the introspective process) as ontologically distinct states that are connected, if all goes well, by the operation of a relatively simple and direct causal mechanism that scans, searches, senses, detects, or keeps tabs on the relevant aspects of the mind. Shaun Nichols and Stephen Stich (2003), for example, argue for the existence of a "Monitoring Mechanism" that detects our beliefs and desires and creates new beliefs about the detected beliefs and desires. David Armstrong describes introspection as a "self-scanning process in the brain" (1968, p. 324) and emphasizes the ontological distinctness of the target state and the introspective apprehension of it (1963, p. 422-423). Versions of the introspective self-detection view go back at least to John Locke, who described 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). Such accounts are often very sketchy about the details – perhaps partly because the cognitive process involved is assumed to be too simple to admit of analysis into detail, at least using the tools of philosophy. (Goldman 2006 is a partial exception to this trend. For a more detailed review of these accounts, and also of accounts rejecting introspective self-detection, see Schwitzgebel 2010.)

Undeniably, there are methods for generating true self-ascriptions of current mental states without introspective self-detection of that sort. I will describe five such methods now. Advocates of self-detection models can acknowledge the existence of such methods, but they will deny that introspection is a matter of deploying them. Other theorists – those who eschew

introspective self-detection approaches – emphasize such alternative methods, suggesting that much or even all of our self-knowledge arises by these other means.

(A.) *Observing the body or the outside world.* I look in a mirror, see my scowling face, and thereby see that I am angry. I observe the output of a brain-imaging device, see that my visual cortex is active, and on that basis conclude that I am having visual experience. I hear myself say, "I think Ohno needs to make his move now" and on the basis of my perception of what I have just said learn that I think Ohno needs to make his move now.

(B.) *Inference.* I notice that I tend to avoid crossing paths with Joan in the hall, that I quickly skip over her emails, and that I rarely chat with her at faculty meetings. I conclude that I must not like her. I read Freud and conclude that I must want to kill my father. I read David Chalmers (1996) and conclude that I have mental properties distinct from my physical properties. (For some reason this gives me great comfort.) Thus, I know about my mind inferentially. "Inference" is intended here to refer to a person-level process and not to subpersonal "inferences" of the sort postulated by psychologists, for example, when they state that the early visual system can infer shape from shading. (Although I find it difficult to draw a sharp line between person-level and subpersonal inference – do spontaneous racist judgments involve inference in the person-level sense? – there appear to be clear cases on both sides; and no argument in this essay, I hope, will collapse with the adjudication of the hard cases.)

Methods (A) and (B) blend into each other to the extent perception is inferential and inference is perceptually grounded. Also, although philosophers dispute the relationship between testimony, perception, and inference (Adler 2006), let's include with (A) and (B) self-knowledge by testimony (e.g., my psychiatrist says that I fear success and I take her word for it).

Occasionally, psychologists speak as though self-knowledge is mostly a matter of applying methods such as (A) and (B) (Bem 1967; Gopnik 1993a&b).

(C.) *Self-attributive expression.* I drop a hammer on my toe and shout "that hurts!" as the result of a linguified version of whatever process normally generates such non-self-attributive outbursts as a cry or an "ow!" Since the latter sorts of utterance presumably emerge spontaneously without a prior self-scanning procedure involving the introspective detection of pain, so also on this model the first sort of utterance can emerge spontaneously without any prior introspective self-scanning. Similarly, perhaps, I can unreflectively burst out with utterances like "don't wanna!" or "I hate you!" One might doubt, especially, that two-year-olds introspect before saying such things. (See Wittgenstein 1953; Bar-On 2004; Gordon 2007.)

(D.) *Self-fulfillment.* I can intentionally create self-fulfilling self-attributions – that is, attributions whose truth conditions are a subset of their existence conditions. I can think to myself, "I'm thinking of a wombat", or even simply "I'm thinking". The self-referential structure of such thoughts ensures that if I succeed in thinking them, they are true. Descartes, of course, famously discusses such self-fulfilling thoughts in his *Meditations* (1641/1984; see also Hintikka 1962; Burge 1988, 1996; Heil 1988). It may also be possible somehow to include a conscious experience within a judgment about the existence of that experience, as in "I am having a visual experience of [this] color" or "I am visually experiencing [this]" or "I am experiencing _____", where the "this" or the blank is filled with the target experience itself in a self-fulfilling or partly self-fulfilling way (Gertler 2001; Papineau 2002; Chalmers 2003; Horgan and Kriegel 2007). Or the disposition to self-attribute an attitude may be partly constitutive of having that attitude (Shoemaker 1996). Such self-attributions needn't involve any sort of scanning or self-detection. For example, I might start the wombat-thought with the plan simply to say "I am thinking of..."

and then utter any word whatsoever as the complement. By uttering the word "wombat", I make it true that I am thinking of a wombat (at least on a liberal view of what counts as "thinking of" something); I do not detect the prior existence of a wombat-thought.

(E.) *Self-shaping*. A seventeen-year-old boy is out on his first date ever. He blurts out, just to impress his date, that he is the kind of guy who buys women flowers. At the same time, he successfully resolves to become the kind of guy who buys women flowers, even though he was not that kind of guy before. Or: I say I am imagining my mother's face, making that statement true in the course of self-attributing it by intentionally imagining my mother's face. Self-shaping differs from self-fulfillment in this way: Self-shaping self-ascriptions are not automatically made true by the self-ascriptive act but rather require some further activity, processing, or resolution on the part of the self-ascriber. (McGeer 1996, 2008; and Moran 2001 offer accounts of self-knowledge that emphasize self-shaping.)

Transparency procedures for self-knowledge, which have recently received considerable attention (e.g., Harman 1990; Dretske 1995; Tye 2000; Moran 2001; Kind 2003; Siewert 2004; Byrne 2005; Gordon 2007), emphasize that self-knowledge often involves looking out at the world rather than scanning or detecting something interior. Transparency approaches are not distinct from the five approaches described above; rather, they usually involve one or more of these approaches: Looking out at the world issues in a self-attributive self-expression, or is the basis of an inference, or is a self-shaping activity, or creates a self-fulfilling judgment. However, some transparency approaches (perhaps Byrne 2005, depending on how his view is interpreted) may also or instead involve something closer to relatively direct sensitivity, in the sense about to be described. (More on that topic in Section 8.)

4. Relatively Direct Sensitivity.

Could the five procedures described in Section 3 (that is, Methods A-E), or near relatives of them, account for everything that we might have thought we knew by introspection? In my view, they come much closer to doing so than most advocates of introspective self-detection would probably be comfortable acknowledging. As the advocates of those five methods have correctly emphasized, such methods play a very large role in generating judgments about our current states of mind. However, I don't think that those five methods exhaust the story; something that is broadly speaking like introspective self-detection must also occur. I call this other thing *relatively direct sensitivity*. Before I clarify the meaning of that phrase, I offer two plausibility arguments against the view that the five methods above exhaust the sources of introspective self-knowledge.

First, it seems that we sometimes know our very recently past mental states by means other than the five above. The clearest example is when we chase back a stream of thought. I'm driving on the freeway and find myself thinking about fixed-gear bicycles. Wondering how I came to think of that, I seem to recall that that thought had been preceded by a thought about my son's former third grade teacher, who is the only person I know to have such a bicycle. And it seems to me that that thought had been preceded by my thinking about needing to set up an appointment with my dentist, who practices near the school where my son went to third grade. And that thought, in turn, had been preceded by my noticing that one of my teeth was bothered by the heat of my coffee. Let's assume that I'm right about what my stream of thought was and also that I wasn't introspecting those first thoughts at the time they occurred. I can't know this recently past stream of thoughts by self-attributive expression, self-fulfillment, or self-shaping

(Methods C, D, and E), since those methods can only support the ascription of current or future mental states, not past ones. Nor do perception and inference (methods A and B) seem to ground my knowledge. They might help me somewhat in recovering my stream of thoughts – maybe I still feel the heat of the coffee, even if it is no longer painful, and maybe the evident associative connections between my thoughts provide some inferential scaffolding. But for the most part, without the help of some apparently more direct knowledge, environmental and psychological evidence appear to be inadequate for the reconstruction of any particular stream of thought. We don't, it seems plausible to suppose, generally know our stream of recent thoughts by inferring what they must have been, given the current environment.

But is such knowledge introspective? Maybe it's better to say that it's immediately retrospective. Still, it follows that some interior causal processes not involving Methods A-E can help connect our mental states with our judgments about them. It would be odd if connective processes other than Methods A-E only operated in the retrospective case and no similar process could help connect our mental states with our concurrent judgments about them. In the concurrent case, the presence of such processes might be harder to discern because those processes will, it seems, almost inevitably become entangled with the mechanisms of self-attributive expression, self-fulfillment, and self-shaping; that's why it is valuable to focus on the immediately retrospective case. And immediately retrospective memory of sense experience and emotions might be hard to disentangle from memory of outwardly perceived events and bodily states; that's why it is helpful to focus on immediate retrospection of thoughts or imagery. But once we acknowledge the existence of such connective processes in at least a limited range of cases, there seems no reason not to generalize: In the concurrent case too, including for sensory experience and emotion, some supportive interior causal connections will exist that are not

exhausted by (even if they are also entangled with) methods A-E. I acknowledge that the evidence I've offered doesn't force the conclusion, which is why I offer it only as a plausibility argument.

My second plausibility argument is just this: Since the brain is massively interconnected, it would be odd if all our judgments about our currently ongoing mental states had to derive exclusively from processes like outward perception, (person-level) inference, self-attributive expression, self-fulfillment, and self-shaping. If there is any evolutionary or developmental advantage in accurate introspective judgments (admittedly, the advantage may be slight), then it seems that the brain ought to be able to recruit relatively direct causal streams of influence between the target mental states and the introspective judgments about them (and to do so even when the judgment does not literally contain or suffice for the target mental state). The burden of proof, it would seem, should fall on someone who denies the possibility of such interconnections, except as mediated by perception, inference, and the rest.

It doesn't follow from these two plausibility arguments that there is any dedicated or functionally isolated mechanism or scanning procedure that connects the target states with the introspective judgments about them. The arguments only support a more modest thesis about the relationship between the targets and the judgments – that the latter are sometimes, as I will say, relatively directly sensitive to the former.

One process is *causally sensitive* to another, in the intended sense, if the two processes are ontologically independent (one, for example, is not a part of the other) and the first differs contingently upon differences in the second due to a causal link between them. So, for example, the growth of a homeowner's lawn is causally sensitive to the placement of fertilizer on it because that growth varies depending upon the placement of the fertilizer, due to a causal

relationship between fertilizer placement and lawn growth. In contrast, the growth of a homeowner's lawn is normally not causally sensitive (I assume) to the commercial success or failure of the latest Lady Gaga video. The growth of a homeowner's lawn is also causally sensitive to the owner's intentions regarding the lawn. The growth patterns of lawns vary with the varying intentions of the homeowners via such intermediary processes as the owner's use of fertilizer or not, mowing or not, etc. However, the growth of the lawn is not *directly* sensitive to the intentions of the homeowner: To affect the lawn, those intentions must operate through intermediary processes. From a certain point of view, even the application of fertilizer does not have a direct effect on the lawn but must operate via certain intermediary chemical processes. This thought can perhaps be pursued down to the quantum level. For practical purposes, then, directness or immediacy is relative to a point of view or set of interests. For the microbiologist, application of fertilizer does not have a direct effect; for the homeowner the relationship is direct enough. So let's say it's *relatively* direct.

One important qualification: In normative cases (and maybe in some non-normative cases), usage of the term "sensitivity" tends to imply a certain *type* of contingency. If the husband always acts to frustrate his wife's desires, we would not ordinarily say he is sensitive to them, although his actions do vary contingently upon his wife's desires. A thermometer is not, in this normative sense, sensitive to very subtle changes in temperature unless it responds *appropriately* to such changes. Likewise for the contingencies involved in introspective sensitivity: If the presence of pain always causes you to judge falsely that you are visually imagining the Egyptian sphinx, then the judgment is in some sense causally sensitive to the pain, but not in the normative sense of "sensitive" intended here. There must be some tendency – at least imperfectly, when conditions are right (but not only when conditions are artificially

supportive), and in broad strokes – for the target mental states to promote judgments that the target mental states are present; the causal relationship must be an appropriate one.

I suggest, then, that our judgments about our ongoing or very recently past mental states are sometimes relatively directly sensitive to those mental states – sensitive in a way that does not involve automatic self-fulfillment or ontological dependence, and sensitive in a way that is not mediated by substantial intermediary processes such as outward perception of the world, person-level inference, the shaping of one's mental states to conform to one's judgments about them, the mechanisms of self-expression, or various other more far-fetched methods (e.g., having a genius neuroscientist read one's mind and then implant the judgment).

I confess that I'm not entirely happy with this characterization of the close causal relationship that sometimes exists between our mental states and our judgments about them. My characterization seems imprecise at important points, especially regarding the normative requirement and regarding what counts as an intermediary process (perhaps particularly in the case of inference). My main idea is this: There can be a close causal, non-ontological connection between judgments about our current or very recently past mental states and those states themselves – a connection not reducible to methods like Methods A-E of Section 3, a connection that also helps to explain whatever accuracy we have in our introspective judgments, and a connection that someone can endorse who is nervous about the idea of "monitoring" in any robust sense of that term. Hopefully, little of what follows hangs on the details of the present section. Hopefully, also, the examples in the remainder of this essay will further flesh out and clarify the abstract formulation of this section, giving my overall approach to introspection a clear enough profile to evaluate.

5. Self-Knowledge of Sensory Experience, Emotion, and Imagery.

Consider, then, three possible models of our knowledge of our sensory experience, our emotion, and our imagery:

Detection After: The target state occurs and then a separate detection process is launched that generates a judgment either about that (now very recently past) state or, alternatively, about the current manifestation of that state on the assumption that it hasn't relevantly changed in the intervening time.

No Direct Sensitivity: We know about the target state but not in a way that meets the criteria (described in Section 4) for relatively direct sensitivity – for example, by detecting some outward object or bodily state and then inferring the target state on a basis of a self-attributive theory or through some act of self-referential self-fulfillment (if that is possible for states of this sort).

Entanglement: We know about the target state in a way that reflects at least a certain amount of relatively direct sensitivity to that state – however, not by means of a separate detection process that is launched afterwards but rather through a temporally extended process that includes, as a part, the processes generating the target state.

Most philosophers and psychologists writing on self-knowledge of sensory experience, emotion, and imagery either endorse or assume, explicitly or implicitly, either a Detection After model or a No Direct Sensitivity model. Self-detection accounts of self-knowledge tend to assume Detection After; philosophers who reject self-detection for some other preferred method of self-knowledge tend to assume that their favored process shows No Direct Sensitivity. In this

section, I will suggest that an Entanglement model is sometimes appropriate instead. But my idea is not to promote a *simple* Entanglement model, corresponding to some straightforward process of introspective self-knowledge. Rather, my aim is to display the complex and multi-faceted processes by which self-knowledge is acquired. The processes driving our judgments about our current mental states are, I suggest, pluralistic, opportunistic, and spatially and temporally distributed, tending to recruit a variety of processes including often the target processes themselves. I offer the Entanglement model as one way of displaying the complexity and diversity of the psychological processes of self-knowledge.

Sensory experience. Recall example 1 from Section 2: I look at a tree and reach judgments about its apparent color, about the visual unclarity of the top branches when I focus on the bottom, and about how the green of the leaves in direct sunlight looks in some way the same as and in some way different from the green of the leaves in the shade. On a simplistic version of the No Direct Sensitivity view, I might know that I'm having a visual experience of greenness as I look at the leaves because I know that the leaves are green and that I am looking at the tree in good viewing conditions, and from these facts, plus some general background theories about how vision works both in general and in myself, I infer that I am visually experiencing greenness. I *could* think about it in that way, couldn't I? Presumably I think that way when I reach judgments about the visual experience of the person standing next to me. Let me suggest that we think about our visual experience *partly* in that way in our own case too. The leaves are green. That's a very salient feature of the world, for me, as I consider my visual experience of color in looking at the leaves. If I have inferential grounds to use that fact in support of my judgment about my own experience, why not do so? It would be silly of me – wouldn't it? – to

ignore my knowledge of the color of the leaves and the nature of vision in reaching judgments about my visual experience.

In fact, judgments about sensory experience can easily collapse into judgments about the outside world. If asked, for each of a series of stimuli, to report on one's visual experience of the color of the stimuli, one might first say "green", then "red", then "green again" – eventually slipping into expressing one's judgment about the stimuli themselves, rather than about one's experience of those stimuli, especially if there's no reason to doubt that one's perception is veridical. The two sorts of judgment seem rather difficult to pull apart, though in cases of illusion the distinction matters and the two judgments have different truth conditions. E.B. Titchener (1901-1905, 1912) and E.G. Boring (1921) sternly warned against this kind of slippage from reporting on experience to reporting on the world, calling it "stimulus error" (or "R-error"; see also Schwitzgebel 2005). Let's shift sensory modalities: I taste a burrito, detecting cheese in it, and I say it tastes "cheesy"; but I have no clear idea what "tasting cheesy" involves experientially. Maybe my wife ordered a burrito with cheese but she can't tell if her order was executed correctly, so I give the burrito a try. In saying that it tastes cheesy, I might then be simply, or mostly, reporting non-introspectively on a perceived fact about the outside world: This burrito has cheese in it.

In some cases, a No Direct Sensitivity explanation might be the whole story: Either my judgment, though expressed in literally self-attributive language, really only reflects an opinion about the outside world; or maybe my judgment does in fact pertain to my experience but is inferentially or theoretically derivative only from my knowledge of the outside world plus general knowledge about myself. In most cases, I suspect that these sorts of processes are part of the story – often, perhaps, the most important part. However, we can acknowledge the

importance of such ways of arriving at self-attributions without rejecting the existence of some relatively direct sensitivity also. When I notice the difference between how the tabletop feels to the pads of my fingers versus how it feels to my fingernails, I appear to be reaching a judgment based not entirely on my knowledge of features of the table and my fingers' motion. So also when I notice a gestalt switch in a Necker cube or a duck-rabbit, or when I think about the changes in my visual experience when I push on the corner of one eye, or when I close my eyes and notice the play of light and darkness. In these cases, I'm noticing something about how I am experiencing or consciously representing the table or the cube or the duck-rabbit or the overall structure of the visual field and its inhabitants; I'm noticing and responding relatively directly to features of my mentality, not (or not only) to features of the outside world. And if I can do so in these unusual cases, then it seems likely that I can do so in the more normal case too – for example, with the greenness. I might, as I look at the leaves, at least be open to the possibility of a disconnection between by perceptual experience and my perceptual knowledge of things outside. I might be ready, for example, for an afterimage or floater, recognized as such, to distort my experience of the shape or color of the tree, in which case my judgments about my visual experience of the tree might diverge from my visually-based judgments about the properties of the tree I'm currently seeing. The absence of such an afterimage or floater might obscure the fact that my judgment that I'm having greenish visual experience has two co-operating bases rather than only one. Again I offer considerations of plausibility: Given that I can easily reach judgments about my sensory experience based on my knowledge of the features of what I am perceiving, and given that my judgments about my experience can at least sometimes be relatively directly influenced by features of my experience, including changes in experience that don't match exactly with what is derivable from knowledge of the objects I am perceiving, why

wouldn't I normally avail myself of both sorts of knowledge in reaching judgments about my experience? (If you're inclined to respond to this rhetorical question by asserting that knowledge of the outside world is redundant because our direct knowledge of our stream of experience is vastly more accurate, I would suggest that we don't actually know the stream of experience very well, as I have argued in recent work: Schwitzgebel 2008, forthcoming.)

In attempting to determine what my visual experience is like outside the narrow region of focus on the lower branches of the tree, I deliberately hold my eyes still. I then attempt to discern structures in the upper branches. Based in part on my failure to discern those structures I conclude that my visual experience outside the precise region of focus is, at least in this case, indistinct. Maybe also, if this is different, I notice that the top of the tree looks blurry. Holding my eyes still, attempting to discern the structure of the branches – this bodily control and failed act of perception are, I suggest, an integral part of the process issuing in my judgment that things look indistinct outside a narrow region of focus. Let me lean here a bit on recent work on embodied or enactive perception (e.g., Hurley 1998; Noë 2004; Clark 2008). Defenders of embodied or enactive theories of perception argue that perception is not helpfully seen as the processing of inputs merely acquired as the result of bodily movement and is better seen – perhaps most compellingly in the case of touch – as a temporally extended process integrally involving bodily movement. So also, I want to suggest, it's natural and helpful to think of my holding my eyes still and attempting to reach a perceptual judgment not merely as antecedents of an introspective procedure but rather as integral parts of it. My "introspective" judgment is grounded in a spatially and temporally extended process involving bodily and ordinary perceptual activity. Likewise, seeing the blood welling from my bare foot, remembering that I just dropped a knife on it, the perceptual or quasi-perceptual nociceptive processes involving the

nerves leading from my foot to pain-related brain areas, and my impulse (inhibited or not) to say "ow!" all co-operate in driving my judgment that I am feeling a sharp pain horizontally across the top of my foot.

The picture to which I'm inviting you is this: In reaching judgments about my perceptual experience, I deploy a variety of processes, including perhaps inferential, perceptual, and self-expressive processes, as well as bodily movement. The resulting judgment will hopefully also, if all goes well, exhibit some sort of relatively direct sensitivity to the experience, in the sense that the experience itself has the right sort of causal influence upon the judgment; but whatever processes constitute that relatively direct sensitivity might not be the cognitive processes most influential in determining the final judgment, and they might not be usefully treated as separable from the other processes or capable of acting entirely on their own.

Assessing my current emotional experience. On a Detection After model of the self-knowledge of emotional experience, we typically first experience an emotion and then a separate detection process produces a self-attributive judgment about it. On a No Direct Sensitivity model, our judgments about our emotional experience reflect no direct sensitivity to the emotional experiences themselves. On an Entangled model, our judgments about our emotional experiences typically reflect a direct causal sensitivity to those experiences, but not through a detection process that operates distinct from the target experience.

Assessing one's current emotion and assessing one's current emotional *experience* might be somewhat different activities with somewhat different aims – unless emotions just are experiences of a certain sort. Theories of emotion often treat, as aspects of emotion, the subject's bodily arousal, her appraisal or evaluation of a situation, her behavior or action readiness, or her facial expression (e.g., Schachter and Singer 1962; Scherer 1984; Ekman and

Davidson 1994; Damasio 1999; Prinz 2004b). The relationship between each of these and a person's emotional experience might be complex: Perhaps part of what it is, sometimes, to experience an emotion is to experience some of these things – for example, to have an experience of (or as of) bodily arousal of a certain sort. Or maybe bodily arousal and the rest, or the experience of bodily arousal and the rest, are just regular concomitants of emotion or emotional experience, rather than aspects of emotion or emotional experience. But whatever stand we take on such matters, it seems likely that the assessment of one's own emotion and one's own emotional experience are entangled with each other and with assessment of one's bodily state, facial expression, situation, behavior, and felt impulses to act.

Accordingly, my judgment (in Section 2, example 2) that I feel tense and slightly angry flows, it seems, from multiple sources, including knowledge of my environment (I'm writing a philosophy essay under deadline pressure) and knowledge of my body (my lips are pursed, my throat tense). In another context, my knowledge that I'm afraid of the rattlesnake or (perhaps differently) that I'm *feeling* afraid of it, might derive in part from my general knowledge that rattlesnakes are dangerous and my visual knowledge that one is only three feet away; it might derive in part from my proprioceptive and visual knowledge that I've just flinched, my knowledge that I just felt a tingling surge of what I would call adrenaline, my sense that I have the impulse to run; it might derive in part from my knowledge that I just uttered an expletive, either in inner or outer speech, from an awareness that I'm imagining the snake biting me, from a kind of numb paralysis I feel; I might have an impulse to say to my hiking partner "I'm terrified of that snake", which I do or do not disinhibit. Depending on how emotional experience is ontologically constituted, some parts of that pattern might be tantamount to relatively direct sensitivity to emotional experience or to an aspect of emotional experience, and others might be

partly constitutive; maybe, too, there's something more to emotional experience that's irreducible to such things – a distinctively fearful quale? – and my emotional self-attribution can be relatively directly sensitive to that also.

In thinking about my tenseness in writing this essay, I engaged in a temporally extended bodily process that involved movement – pursing and unpursing my lips – and a sequence of conjecture and theorizing. I thought sequentially about my lips, my chest, my situation and the emotional response that situation is likely to evoke; I reached a tentative judgment, then noticed my lips again, had second thoughts about my judgment, noticed my throat and brow, and suspected that my emotion was worsening in the course of considering it. Had I wished to pursue the matter further, to assess if I really was tense about the deadline, I might have tried focusing on the deadline's imminence to see if my tension increased. Also: Maybe when I began my reflections what was really going on emotionally with me was rather unspecific negative affect, and that affect only became specific anxiety about the deadline as a result of my theoretically-driven self-attribution of such anxiety.

Even these reflections, complex as they are, probably oversimplify the phenomena. Such a chaotic buzz puts pressure, I think, on simple models of self-knowledge.

Imagery. Imagery is so much under our immediate control that concurrent introspective judgments about it seem bound, in most cases, to be supported to some extent by self-shaping – maybe even self-fulfillment if the target image can be a part of the self-attributive judgment about it. When I say that I'm visually imagining my mother's face or hearing the chorus of "Kool Thing" in my head, I am partly making these self-attributions true as I reach them. However, it seems unlikely that I can make just any judgment about my imagistic phenomenology true simply by willing it to be so: If I judge that I am visually imagining the Taj

Mahal with every arch and spire simultaneously well defined, or that I am imagining, simultaneously, the voice, bass line, drums, and both guitars, or that I am visually imagining a triangle that is neither equilateral, isosceles, nor scalene but somehow all and none of these at once (Locke 1690/1975; contra Berkeley 1710/1965), I might be wrong. And hopefully if I am wrong, something in me – some influence from the imagery experience itself – will (even if not dependably) cause me to refrain from the attribution, or cancel it, or at least hesitate and feel uncertain. When I'm trying to determine if I can imagine the lead guitar and vocals at the same time, it seems that I'm not only creating or sustaining the imagery but also checking to see if I have successfully created it as intended.

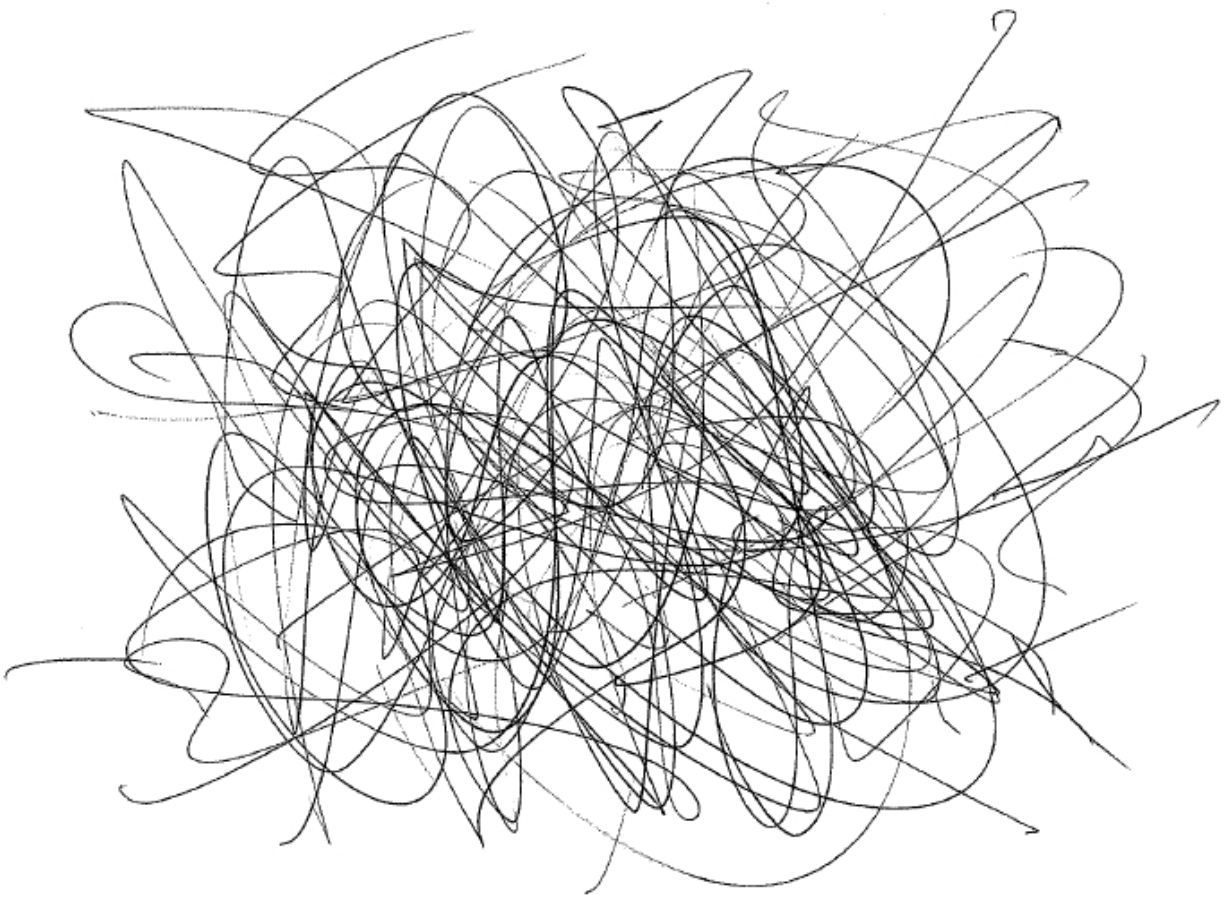
As mentioned in Section 4 above, self-shaping and self-fulfillment cannot explain knowledge of very recently past imagery, and often there is little environmental or inferential basis for judgments about the contents of one's recently past imagery; so this is a case where relatively direct influences from conscious experience to one's judgments about it are most evident. Those influences might be only marginally direct, however, mediated by short-term memory, or iconic memory, or a looping process; or they might be more unmediated, reflecting fading activation or the normal temporal course of a feed-forward causal brain process or partial temporal overlap between processes.

This isn't to say that our judgments about our imagery are especially trustworthy. Judgments about imagery vividness, for example, are suspiciously uncorrelated with behavioral performance (Schwitzgebel 2002, forthcoming), and judgments about other structural features – e.g., flatness, stability, and coloration of parts for which coloration is not a salient feature – seem likely to be guided as much by implicit or explicit psychological theory as by sensitivity to the presence or absence of those features in the target image.

The boxology of self-knowledge. I haven't yet addressed knowledge of attitudes and the example of deciding about the pie; that's for Section 9. I hope, however, that the present examples lend plausibility to a complex Entangled view of our knowledge of our own current conscious experience – a view on which judgments about one's current experience are driven by a wide variety of interacting factors, and on which the processes issuing in such judgments are often spatially and temporally extended, including as a part the target processes themselves or aspects of them.

It's often helpful for cognitive scientists modeling psychological processes to describe the mind's functional architecture using boxes and arrows, with the boxes indicating various functionally discrete processes or systems and the arrows indicating the causal or functional relationships among those discrete processes or systems. Figure 1 expresses my view of self-knowledge, using the "boxology" of cognitive science. The model in that figure may be contrasted, for example, with the boxological models on pages 162 and 165 of Nichols and Stich 2003, which feature tidy arrows in and out of the Belief Box, through a Monitoring Mechanism, a Percept-to-Belief Mediator, and a Theory of Mind Information store. You might also notice a resemblance between my model in Figure 1 and recent boxological models of visual processing, if the latter are squinted at.

Figure 1: The boxology of self-knowledge



6. Introspection, What.

In the previous sections, I have mostly spoken of "self-knowledge" or "self-attribution" rather than introspection. That's because to call all of the various processes "introspective" seems a stretch, maybe even an abuse of the term. In this section, I describe criteria that distinguish what I would call introspective processes from non-introspective ones. Preliminary caveats: (i.) I doubt sharp lines can be drawn through the tangle. (ii.) Even if a process, considered on its own, is not introspective by the criteria I'm about to articulate, it may be embedded in a larger process that is introspective (as with many of the sub-processes in the previous section). (iii.) I have been writing and re-writing this essay in various versions since 1999, every year or two tweaking my formal definition of introspection. Probably, in another

year or two I'll think the definition needs further tweaking. I offer it in the spirit of a rough sketch.

Here, then, is my definition of introspection: To introspect is to dedicate central cognitive resources – that is, attention – to the goal of reaching a judgment about one's current, or very recently past, conscious experience, and to do so in a way that aims to exhibit direct sensitivity to that experience. Now, let's unpack.

Can we attend to experience? Several prominent contemporary philosophers have said that it's impossible to attend to one's sensory experiences – for example, Gilbert Harman (1990), Fred Dretske (1995), and Sydney Shoemaker (1996). (For critical discussion, see Kind 2003; Siewert 2004.) If they are right, introspection of sensory experience – in my sense of "introspection" – would be impossible. Harman's discussion 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 (1990, p. 667).

Similarly, Dretske writes:

> If one is asked to introspect one's current gustatory experience – "Tell us, if you can, exactly how the wine tastes" – one finds oneself attending, not to one's experience of the

> wine, but to the wine itself (or perhaps the tongue or palette). There seems to be no other relevant place to direct one's attention (1995, p. 62).

Such claims are plausible, I think, only on an impoverished and overly sensory conception of attention.

To motivate a less impoverished conception of attention, let me present a few non-sensory examples of attention. We can attend, it seems clear, not only to material objects and spatial regions but also to things like arithmetic problems and word puzzles entertained entirely in the mind, to the task of figuring out the best way to grandma's house during rush hour, to the task of pushing oneself physically on an exercise machine, or to one or another properties of a single object in a single spatiotemporal region (e.g., an apple's color vs. its shape vs. how it reflects the overhead lights). Sophisticated general accounts of attention, from John Dewey (1886) to Harold Pashler (1998) thus tend to characterize attention not in terms of selection from among physical objects or regions but rather as selection among, or dedication of resources to, cognitive tasks or processes. In attention, our very limited central cognitive resources – at least roughly speaking, and perhaps precisely speaking, the processes of conscious thought – are either devoted to a single task or split among some few. Normally, there is some cognitive goal or aim toward which these resources are devoted (figuring out the best way to grandma's house, obtaining precise knowledge of the apple's shape, etc.) – an aim that the attending person may or may not acknowledge or endorse. (When attention is captured exogenously by a surprising stimulus, for example, no endorsement of the aim of detecting the properties of that stimulus is necessary.)

Clearly, in this sense of attention, one can attend to one's experiences – that is, one can devote central cognitive resources with the goal or aim of determining their properties. One can

think about them. (This would be so, even on the view that the only way we can think about our experiences is by inferring what they must be, given the state of the outside world.) Equally clearly, no one should be tempted to think that we can attend in a *sensory* way to our sensory experiences, if that implies taking our sensory experiences as ordinary objects of perception. We don't see our visual experience, after all; no light reflects from it into our eyes. So what is it that we are supposed to conclude from Harman's and Dretske's examples? In what sense can't we attend to our experiences? It is now mysterious to me. Maybe the view is that the only kind of attention we can devote to sensory experiences is an inferential, abstract, intellectual attention, like the attention we give to a word puzzle or philosophical problem? If so, that claim is neither clear in their work nor well supported by their evidence.

Here's why I think their examples have a superficial pull: If we think of attention – as we too often do – as something that only takes material objects or spatial regions as its objects, then it's natural to suppose that attention to a tree or to a mouthful of wine would compete with attention to one's experience of the tree or the wine. After all, the tree and wine are in one place and one's experience of them (presumably) in another; thus it seems that we would have to withdraw attention from the one to shift it to the other, and it doesn't seem like we do that. But goals nest differently from objects and regions. We can achieve one goal in part by achieving another. As Dretske himself emphasizes (in different terms), one's goal of determining the properties of one's visual experience might contain as a subgoal one's determining the properties of the tree. This follows naturally from the Entangled model of self-knowledge presented in Section 5. Thus, on the goal model of attention, as opposed to the material objects / spatial regions model, dedicating attention to experience does not imply withdrawing it from the objects of experience. (See also Siewert 2004.) For a non-introspective example of such nesting,

consider the goal of determining the warp of a window by noticing how things seen through it seem to noodle around when you move your head. What we should learn from reflection on Harman's and Dretske's examples is not that experience can't be attended to; rather, it's that knowledge of experience is entangled with knowledge of the world.

Does introspection necessarily involve attention? Introspection, as I conceive it, is an attention-hogging process. It's not something we do all day. Rather, it involves turning one's thoughts to the task of discerning features of one's own mind, specifically features of one's stream of conscious experience. In characterizing introspection this way, I am, I think, picking out a particular sort of interesting activity relevant to the projects of philosophy, psychology, and consciousness studies.

Must introspection aim at reaching a judgment? Philosophers and psychologists have often characterized the product of introspection as knowledge, awareness, belief, or judgment. I opt for the last of these: "Knowledge" assumes correctness (one can't know what is false), and I think introspection can err (see esp. Schwitzgebel 2008, forthcoming). The term "awareness" I find unclear and confusing, partly because it seems to waver been a phenomenal sense (meaning "conscious experience") and an epistemic sense (meaning, roughly, "knowing about something"); also, like "knowledge", it assumes correctness. "Belief" is tempting because it is such common coin in contemporary analytic philosophy and does not assume correctness. However, I find the term too dispositional to serve as the product of introspection: One believes something if one has a certain overall functional or dispositional structure. No particular episode of "believing" must transpire. A "judgment", on the other hand, is a particular episode of endorsing something or taking it as true; it is, perhaps, part of the occurrent or episodic side of belief. That episode – that judgment, or endorsement, or act of categorization, or taking as true –

is the normal product of introspective activity. By "judgment" I don't mean anything particularly highfalutin or intellectual – an inarticulate appreciation qualifies. We introspect and arrive at some momentary assessment, not necessarily verbalizable, of our conscious experience.

But I say that introspection need only *aim* at reaching a judgment: It seems plausible that an introspective process can fail or be aborted and thus issue in no judgment. Also introspection *must* aim at reaching a judgment (at least in the sense in which cognitive processes can aim at goals independently of our person-level aims): If some attention-consuming process results only incidentally in a judgment about current conscious experience, without that having been part of the goal of the process, that process doesn't seem worth calling introspective.

Many self-attributive utterances will not qualify as introspective by this criterion. If I say "I don't think the Lakers will win this game" meaning only to be making a statement, in a hedged way, about the Lakers' chances, then although the statement may literally be a true self-attribution, the process issuing in that statement was not a process aimed at issuing in a judgment about my own experience. The same is true if I say "I'd like the red one" just as a polite way of saying "give me the red one", or if I say "that looks like a hard shot" as a way of saying it probably is a hard shot, or if I say "I'm fine" merely as a way of acknowledging someone's how-are-you. Such processes, I'd suggest, *can* be launched with the aim of producing a self-attributive utterance that is tantamount to a judgment about one's own mind, but they need not be launched with that aim. Probably, also, the aims in launching such processes can be multiple or indeterminate, contributing to the blurry lines between introspective and non-introspective processes.

Is the introspective target current, very recently past, or both? Presumably, I don't know about yesterday's emotions and imagery by introspecting them. I know about them through

memory. (Maybe I can introspect memories, but then it is the memories that are introspected, not yesterday's mental states.) Introspection, as normally conceived, transpires in a narrow temporal window. It's natural to think that introspection concerns current mental states only, since introspective judgments are typically framed in the present tense. But there are two reasons to think that introspection might also encompass the very recent past. First, if the introspective process is a causal process that elapses over time, the target process might be over and gone by the time the introspective process is complete, or at least the relevant time-slice of the target process might be over and gone. If the introspective judgment nonetheless attributes a mental state indexed to the time at which the judgment is completed, it must work on the implicit assumption that the target state and its relevant features will have persisted (Armstrong 1963). A more cautious introspective judgment might, then, work by essentially the same process but refer to the very recent past. Second, if cognitive processes are distributed across time, a sharp distinction between the exact present and the very recent past may be misleading. It's an open question how far in the past a mental event must be before the processes by which we arrive at judgments about it are sufficiently different from the processes by which we arrive at judgments about current mental events to warrant regarding the former as necessarily non-introspective. (Addressing this last issue might be one way to draw the boundary of the "specious present"; James 1890/1981; Varela 1999; Kelly 2005.)

Aiming at relatively direct sensitivity. Introspection, as I have defined it, aims at relatively direct sensitivity. I add this condition to exclude cases like the following: Look in the mirror. See that you are wearing a hat. Apply a theory according to which whenever something is touching one's skin one has tactile experience. On the basis solely of that theory and the visual evidence, conclude that you are having tactile experience. Or: Listen to what you are

saying. Assume that you believe what you are saying. Ascribe beliefs to yourself wholly on that basis. You might try to engage in such processes and fail, of course. You might not be able to reach a conclusion wholly on such indirect and external bases, perhaps because you launch an introspective process, contra your intentions. Or you might successfully avoid introspecting but your conclusions might be directly sensitive to your target states anyway (e.g., by some glitch). Inversely, an introspective process might aim for direct sensitivity and fail to achieve it – for example, if one's introspective judgments about one's experience are massively distorted by one's theories about experience. I suggest that what matters, then, is whether the process aims at being relatively directly sensitive. Direct sensitivity is an epistemic virtue of introspection, among its constitutive aims, but achieving that virtue is not guaranteed.

This condition also excludes cases where the relationship between the judgment and the target state is constitutive rather than causal – cases like "I'm thinking of a wombat" which are self-fulfillingly true whether they arise from a process that seems introspective or from any other process whatsoever including freak accident. When judgments of this form are introspective, it will be because they are sufficiently temporally extended that there is a directly sensitive causal relationship between the thought at time 1 and the judgment about it at time 2 (though the temporal reference of the self-attribution may be vague). Possibly also the subject applies concepts to the embedded contents in a way driven by a causal process rather than simple self-fulfillment. In such cases – and in other cases, too, perhaps especially with emotion – self-attributions may be *partly* grounded in a causal introspective relationship and at the same time partly self-fulfilling.

I would like to be able to say more about what it is for a psychological process to have a goal or aim, but that's a hornet's nest, so I will have to leave that aspect of this criterion sketchy.

Maybe the examples help. I don't want to be so liberal about "aiming" that any arbitrary goal can be turned into an aim of a cognitive process regardless of the ordinary function of processes of that type. The main idea in focusing on aiming at direct sensitivity is to rule out two types of cognitive processes – those which, if successful, result in a self-attributive judgment that is shielded from influences other than indirect ones, and those that are merely self-fulfilling without a causal introspective story behind them – and to rule in cognitive processes that are in all respects like canonically introspective processes except that because of distorting influences on the outcome they fail to exhibit the proper sensitivity.

Must the target of introspection be conscious experience? Everyone will, I think, grant that the targets of introspection are always, in the first instance, one's own mental states and not, except derivatively, non-mental facts or facts about other people's mental states. But not all mental states are conscious. Personality traits, for example – extraversion or kindness – seem not to be conscious experiences but rather dispositions to act and react in certain ways. Although in ordinary language usage people sometimes say that they are introspecting when they are thinking about their personality traits, few philosophers think of personality traits as directly introspectible. It seems that we know about personality traits differently from how we know about current imagery, emotional experience, and sensory experience. It seems that we know personality traits only by relatively indirect means, such as inference from patterns of remembered behavior or by drawing conclusions on the basis of incipient emotions that arise when various possibilities are imagined. Non-conscious cognitive processes – the mechanisms of early visual processing, for example, or the desire to kill my father that my psychoanalyst finally convinces me, on theoretical grounds, that I must have – also appear to be known only relatively indirectly, not introspectively. The paradigmatic targets of introspection are all states

that at least *can* be conscious, in the sense of having a phenomenal or qualitative character, a "what-it's-like-ness": that is, emotions, imagery, sensory experience – and maybe attitudes, though attitudes are a complicated case, which I'll address in Section 9. Bracketing attitudes until that section, it seems that I can know my mental states introspectively only insofar as they are consciously experienced, and then only the aspects of them that are part of my conscious experience. Maybe there is a deep structural reason for this fact; but I draw it only as an inductive conclusion. The only target mental states to which it seems right to say I have introspective access are conscious experiences.

Let me conclude this section by repeating my definition, the motivation and meaning of which is hopefully now clearer: To introspect is to dedicate central cognitive resources – that is, attention – to the goal of reaching a judgment about one's current, or very recently past, conscious experience, and to do so in a way that aims to exhibit direct sensitivity to that experience.

7. Introspective Judgment vs. Attunement to Experience.

People might have some non-introspective *attunement* to the stream of conscious experience through processes that don't require attention or issue in judgments. Such processes seem to me different enough in kind from paradigmatic introspective processes to merit a different label.

To explain what I mean by "attunement" here, let me start with examples not involving self-knowledge. You are at a cocktail party and feel the urge to move on to a new conversation. You have this feeling because you are picking up subtle cues that your conversation partner is bored. Her eyes are wandering a bit, perhaps, or she is exhibiting micro-expressions of disgust,

or she is shifting her weight between her feet a little too often. You may not know that she's bored and you may have no conscious appreciation of the subtle cues. But still you are sensitive to those cues and you respond appropriately with the impulse to move on. Or: You raise your hand to catch a fly ball. It lands in the netting of your mitt, and you know that it does, though you don't see the ball go in because you are already looking at second base. How do you know you caught the ball? By the increased weight of your mitt? By the reflexive clenching of your wrist? By the jostling back of your arm? Let's suppose that it's mainly by the reflexive clenching: If a physiologist could induce that clenching though you miss the ball, you'd think you had caught it; if she could suppress the clenching though you caught the ball, you'd think you hadn't caught it; and parallel remarks are not true, or as true, for the other potential means of knowing. You are, then, attuned to the clenching of your wrist as a sign of ball-catching. But your conscious theory of how you know, if you have a conscious theory, might be just a bunch of nonsensical fluff.

Returning to self-knowledge: We might be non-introspectively attuned to our stream of experience, even when we devote no introspective attention to it. We might spontaneously track and respond appropriately to our sensory experience, imagery, and emotional experience without deploying introspective attention. Abnormalities in experience or intense experiences might then attract our introspective attention. Or we might skillfully respond to variations in emotional experience by self-soothing strategies or by choosing activities that suit the emotion. Likewise, we might be responsive to pain-experiences without having to attend to them. If we can non-introspectively expressively self-attribute conscious experiences (Section 3, Method C), such expressions may also reflect attunement. One way to read "veil of perception" theories of perception (e.g., Locke 1690/1975) would involve attributing to people an attunement-like

knowledge of sensory experience as the psychological basis and epistemic grounds of knowledge of the outside objects represented in that experience. The details of how and when we are attuned to our conscious experience, without introspective judgment, will turn on issues that I can't begin to explore here; but I want to flag the probability of this sort of non-introspective epistemic relationship to one's own experience. (Some other philosophers who distinguish attentive from non-attentive self-knowledge include Brentano 1874/1973; Dainton 2000; Rosenthal 2005; Kriegel 2009.)

8. Shifting Directly from a Judgment.

Alex Byrne (2005) posits "epistemic rules" of the following sort: If conditions C obtain, believe that P. For example: If the doorbell rings, believe that there is someone at the door (a good rule, generally speaking). Or: If *The Weekly World News* reports that P, believe that P (not such a good rule, given the journal's emphasis on Bigfoot and aliens). Byrne says little about the psychological mechanisms required for following such rules, but he seems to regard them as relatively bare. Byrne then suggests that the following is a good epistemic rule:

If P, believe that you believe that P.

Indeed, that is a fairly good rule. It will tend to generate, it seems, true self-attributions. Byrne, in fact, claims that the rule is "self-verifying", with the truth of the resulting belief "guaranteed". He can say this, I think, because he does not distinguish as I would between belief and judgment. (I think the two can come apart, for example, when someone who is racist in virtually all her reactions and implicit assumptions nonetheless sincerely endorses the claim that all the races are intellectually equal. This would be judgment without belief, or at least with only marginal or in-between belief. See Schwitzgebel submitted.) On Byrne's view, applying the rule requires that

you believe that P is true, so the conditions of the rule's execution are sufficient for the truth of the self-attribution that is the rule's product. Even if, as I think, judgment and belief do sometimes come apart, there is a close enough link between judging that P and believing it that the resulting self-attribution will often be true. Dretske (1995) and Tye (2000) offer similar accounts that involve shifting from representing x as F to representing oneself as representing x as F.

I see two ways to execute rules of this sort. Adapting Byrne's language to my own preferred language of judgment, I will call the first way of following the rule *dual-purpose judgment*. First I judge that P. Then, a moment later (or perhaps simultaneously but using P as the grounds?), I judge *both* that P and that I believe that P, in such a way that the judgment that P is embedded reflexively within the self-attributive judgment. Such judgments have what Tyler Burge (1996) calls "herewith"-like reflexivity (as in, "I judge, herewith, that there are physical entities", although the "herewith" can be, as it were, silent or implied). The judgment I arrive at by following the rule is thus simultaneously both self-attributive and directed at the outside world. It is dual-purpose. It's self-fulfilling, and thus not a matter of introspective sensitivity between ontologically distinct states or processes.

A second way to interpret Byrne's strategy is what I will call *shifting directly from judgment*. This method is best appreciated by considering not belief but desire. Byrne suggests that in self-attributing desire we can also employ an epistemic rule: If ψ-ing is a better option than χ-ing, believe [or judge] that you prefer to ψ than to χ. Such a rule, Byrne says, is *not* self-verifying, since belief [or judgment] that one option is better than another only tends, he says, *typically* to line up with desiring the better option. Because of that typical relationship, the epistemic rule is fairly good. One can also imagine somewhat dodgier rules for other self-

attributions, which still might serve as rules of thumb: If x is dangerous, believe (or judge) that you are afraid of x. If x is low-class, believe (or judge) that you don't like it. If a green bottle is in your field of view, believe (or judge) that it visually appears to you as if a green bottle is in your field of view. And so on. On a theory like mine where belief and judgment can come apart, shifting directly from a judgment *that P* to a judgment *that one believes that P* is not self-fulfilling but merely a fairly good strategy, just as these other methods are only fairly good. And we probably do employ a wide variety of such rules of thumb; the methods are only psychologically implausible if regarded as operating solo rather than in simultaneous competition and co-operation with other methods. (Is such shifting "inference"? I'm inclined to think not – or at least not in the relevant sense – unless the epistemic rule is deliberately applied. But spin that out however you like.)

Such cases of shifting directly from judgment are non-introspective because such processes do not aim to exhibit direct sensitivity to the target state. Instead, the direct sensitivity is to a judgment that is assumed to covary, reliably enough, with the self-attributed target state. There is no direct sensitivity in such a case, I should say, unless the self-attributed target state *is* actually the precipitating judgment. In that last case – shifting directly from a judgment that P to the judgment that you are judging that P – is the process is introspective, then, on my account? Well, that depends on whether judgments are conscious experiences and thus appropriate targets for introspection. That is the topic of the next section.

9. Self-Knowledge of Attitudes.

The previous section concluded with two open questions: whether shifting directly from a judgment to a self-attribution of a judgment is introspective, and, relatedly, whether judgments

are conscious experiences and thus, on my model, appropriate targets for introspection. We need to address the second of these issues before we can address the first. And then we can finally return to the example (from Section 2) of deciding about the pie.

The literature on the consciousness, or not, of attitudes is complicated (e.g., Siewert 1998; Pitt 2004; Robinson 2005; Prinz 2007; Carruthers 2009, 2010). I wish to take no stand on it, except to say this: *If* attitudes are the kinds of things that can be consciously experienced in the sense that they have a phenomenal character, then we can introspect them. (Of course attitudes can be conscious in the broader sense that we can endorse them while self-attributing them, but that's not the issue.) And if attitudes cannot be consciously experienced – if they have no phenomenal character but are instead only associated with particular types of consciousness (for example, with uttering sentences in inner speech) – then we cannot introspect them.

That view fits my account of introspection, of course, which is motivated independently of facts about the consciousness and introspectibility of attitudes. And it has, I think, at least some prima facie plausibility. Warm yourself up, if you can, to the idea that attitudes can be conscious in the sense of having a phenomenal character. For example, maybe intensely desiring fame has a particular phenomenal character. If it seems so to you, then I venture to assert that it should also seem plausible to you that you could be introspectively responsive to that phenomenal character and self-attribute an intense desire for fame in part on the basis of that introspective sensitivity. Inversely, try warming yourself up, if you can, to the idea that attitudes have no phenomenal character. All you experience when you seem to feel an intense desire for fame, say, is visual imagery of adulating fans and a certain emotional experience, which are not constitutive of the attitude. On such a view, attitudes are, let's imagine, like character traits, dispositions to act and react in certain ways. I venture to assert that it should seem plausible to

you, on such a view, that you have no relatively direct introspective sensitivity to the attitude. You can only be indirectly sensitive to the attitude by being directly sensitive to the associated imagery and emotion, probably in conjunction with theorizing and other business. Or maybe there could be a close functional link between the target state and the self-attribution, but only a link so similar to the type of link between target state and self-attribution in non-introspective cases (e.g., in direct shifting from a judgment to the self-ascription of desire or, possibly, between the trait of being shy and the impulse to confess shyness) that the mechanism seems best classified as non-introspective. My account of introspection would then generate the right result under these variations. Introspective judgments pertain only to conscious experiences, including attitudes if and only if to have the relevant attitude is to have a conscious experience.

One of the disjuncts in that last paragraph depends on a hunch that I have been trying to do without: the hunch that there's a difference in kind between the processes characteristic of introspective attention to conscious experiences and the processes characteristic of otherwise similar attention-consuming reflections on mental states that are not consciously experienced. I can circumvent reliance on this hunch if conscious experiences are the only targets to which we can launch attention-consuming cognitive processes aiming at relatively direct sensitivity. If so, then – as suggested near the end of Section 6 – the restriction of introspective targets to conscious experiences follows from the remainder of the definition. However, if not – that is, if there are attention-consuming cognitive processes that aim to reach relatively directly sensitive judgments about mental states that are not conscious experiences – then my restriction of the targets of introspection to conscious experiences requires some justification. Attitudes are the candidate targets that most plausibly require special justification for their exclusion. But attitudes will require special justification for their exclusion only if three conditions are met: (1.)

They are not conscious experiences. (2.) Our judgments about them don't embed them or otherwise show an ontological dependence incompatible with direct sensitivity. And (3.) Cognitive processes can in fact aim to be directly sensitive to them, instead of only being directly sensitive to something else nearby (like imagery or like a consciously experienced judgment associated with but not tantamount to the target attitude). If all three conditions can be met – which is not, I think, clear – then to justify the exclusion of attitudes from the allowable targets of introspection I must fall back upon my impression, which I hope that you the reader might also share, that there is something distinctive about our attention-consuming direct sensitivity to our stream of experience that justifies categorizing judgments that aim to rely upon it in a class by themselves. On the other hand, if the three conditions cannot be met, then the restriction of targets to conscious experiences falls naturally out of other, independently justified, features of the definition.

Let's go back, then to the method of shifting directly from the judgment that P to the judgment that I am judging that P. If this is genuinely a case of shifting rather than a case of self-fulfilling dual-purpose reflexivity, then the process qualifies as introspection on my account only if the judgment that I am judging that P is the judgment that I am having a conscious experience of a certain sort, that is, the conscious experience of judging that P (though the experience need not be conceptualized in those terms). The process must also consume attention and aim at relatively direct sensitivity – including not being merely inferential in a way that implies no direct sensitivity.

Finally, deciding about the pie. I don't want to claim that the preference for apple over cherry is, or is not, conscious in the sense of having a particular phenomenal character. However, I do want to emphasize, as in the other examples I have discussed, the complex tangle

of processes that seem to be involved in reaching the self-attributive judgment: I think about pie (that is, the external thing) and the restaurant's quality; I call up memories; I imagine; I am, perhaps, sensitive to certain felt cravings. I look at the menu – maybe it contains pictures, maybe just words – and doing so helps me resolve my preference or discover it (probably a bit of both). Maybe part of my resolving involves generating an impulse to utter "I'd like the apple", which I refrain from inhibiting. Maybe there is some causal pressure to shift directly from judging that the apple is likely to be better to judging that I want it. Thin versions of most of these processes might going on simultaneously, even if it seems to me that my self-knowledge is simple.

10. Other Pluralist Accounts.

At the core of my view of introspection is a pluralism about process. But there must be *some* limits to what counts as introspective, lest any process in the world count as introspective. I have proposed a definition to sketch the limits, but I readily admit that the definition is only rough and turns on some underdeveloped ideas like relatively direct sensitivity and cognitive goals. Recall the analogy with which I began this essay: taking in a science poster. There is no single, dedicated mechanism for taking in a science poster but rather an opportunistic plurality of processes and sub-processes, some of which, considered on their own, might be only tangentially related to the goal of taking in the poster. And yet not just any cognitive process counts as taking in a science poster: Minimally, the process is an attention-consuming one that aims to arrive at judgments about the poster, and to do so in a way that exhibits a certain visual sensitivity to the displayed poster. Analogously, introspection is a pluralistic, opportunistic, attention-consuming process that aims to produce judgments about current or very recently past conscious experience, and to do so in a way that exhibits relatively direct sensitivity to the target experience. The

Entanglement between the processes generating the target state and the processes generating the introspective judgment is not captured by this analogy (perhaps, though, we could imagine that the poster is being created on the fly partly in response to my questions about it?), but Entanglement is only, on my view, a common feature of introspective processes, not a universal feature. I hope I have displayed some of the attractions of this picture of introspection, especially for readers already drawn toward a general view of cognition as spatially and temporally extended and resistant to neat boxology; but I confess to doubting that I have said anything that should sway someone not already pre-disposed toward a view of that sort.

Jesse Prinz (2004a) and Christopher Hill (2009, forthcoming) have also endorsed explicitly pluralistic views of the mechanisms of introspection. However, my account differs in two important ways from theirs. First, they seem to want to count more types of processes as introspective. This is especially true of Prinz, who seems to regard almost any process that involves "accessing" one's mental states as introspective – even just ordinary remembering (accessing the memory store). Second, their pluralism seems mainly to be a between-judgments pluralism: On their view, many different types of processes can issue in introspective judgments, but the processes issuing in any one introspective judgment, or any one type of introspective judgment, are often relatively simple processes. In contrast, I'm inclined to think that most introspective judgments, when not simplified by philosophers into one-line examples, issue from a cognitive confluence of crazy spaghetti.

Acknowledgements:

For helpful comments and discussion special thanks to Dorit Bar-On, Tim Bayne, Chris Hill, Victoria McGeer, Russell Pierce, Charles Siewert, Maja Spener, Aaron Zimmerman. Many others, I'm sure, are unjustly forgotten. I also thank audiences at U.C. San Diego, University of Southern California, Toward a Science of Consciousness, U.C. Riverside, Macquarie, Australian National University, University of Bristol, Oxford, and York University (England).

References:

Adler, Jonathan (2006). Epistemological problems of testimony. *Stanford Encyclopedia of Philosophy.*

Armstrong, David M. (1963). Is introspective knowledge incorrigible? *Philosophical Review, 72,* 417–432.

Armstrong, David M, (1968). *A materialist theory of the mind*. London: Routledge.

Bar-On, Dorit (2004). *Speaking my mind*. Oxford: Oxford.

Bem, Daryl J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74,* 183–200.

Berkeley, George (1710/1965). *A treatise concerning the principles of human knowledge*. In *Principles, Dialogues, and Philosophical Correspondence*, Colin M. Turbayne (ed.). New York: Macmillan.

Boring, Edwin G. (1921). The stimulus-error. *American Journal of Psychology, 32,* 449–471.

Brentano, Franz (1874/1973). *Psychology from an empirical standpoint*, Oskar Kraus and Linda L. McAlister (eds.) and Antos C. Rancurello et al. (trans.). London: Routledge.

Burge, Tyler (1988). Individualism and self-knowledge. *Journal of Philosophy*, 85: 649–663.

Burge, Tyler (1996). Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society, 96,* 91–116.

Byrne, Alex (2005). Introspection. *Philosophical Topics, 33,* 79–104.

Carruthers, Peter (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32,* 121–182.

Carruthers, Peter (2010). Introspection: Divided and partly eliminated. *Philosophy and Phenomenological Research, 80,* 76–111.

Chalmers, David J. (1996). *The conscious mind*. New York: Oxford.

Chalmers, David J. (2003). The content and epistemology of phenomenal belief. In *Consciousness: New philosophical perspectives*, Quentin Smith and Aleksandar Jokic (eds.). Oxford: Oxford.

Clark, Andy (2008). *Supersizing the mind.* Oxford: Oxford.

Damasio, Antonio (1999). *The feeling of what happens.* Orlando: Harcourt.

Dainton, Barry (2000). *Stream of consciousness.* London: Routledge.

Descartes, René (1641/1984). *Meditations on first philosophy*, in *The philosophical writings of Descartes*, vol. 2, John Cottingham, Robert Stoothoff, and Dugald Murdoch (eds.). Cambridge: Cambridge.

Dewey, John (1886). *Psychology.* New York: Harper.

Dretske, Fred (1995). *Naturalizing the mind*. Cambridge, MA: MIT.

Ekman, Paul, and Richard J. Davidson, eds. (1994). *The nature of emotion.* New York: Oxford.

Gertler, Brie (2001). Introspecting phenomenal states. *Philosophy and Phenomenological Research*, *63,* 305–328.

Goldman, Alvin I. (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.

Gopnik, Alison (1993b). Psychopsychology. *Consciousness and Cognition, 2,* 264–280.

Gordon, Robert M. (2007). Ascent routines for propositional attitudes. *Synthese, 159,* 151–165.

Harman, Gilbert (1990). The intrinsic quality of experience. *Philosophical Perspectives, 4,* 31–52.

Heil, John (1988). Privileged access. *Mind, 97,* 238–251.

Hill, Christopher S. (2009). *Consciousness*. Cambridge: Cambridge.

Hill, Christopher S. (forthcoming). How to study introspection. *Journal of Consciousness Studies.*

Hintikka, Jaakko (1962). Cogito ergo sum: Inference or performance? *Philosophical Review, 71,* 3-32.

Horgan, Terence, and Uriah Kriegel (2007). Phenomenal epistemology: What is consciousness that we may know it so well? *Philosophical Issues, 17,* 123–144.

Hurley, Susan L. (1998). *Consciousness in action.* Cambridge, MA: Harvard.

James, William (1890/1981). *The principles of psychology*. Cambridge, MA: Harvard.

Kelly, Sean D. (2005). Temporal awareness. In *Phenomenology and philosophy of mind,* David W. Smith (ed.). Oxford: Clarendon.

Kind, Amy (2003). What's so transparent about transparency? *Philosophical Studies, 115,* 225–244.

Kriegel, Uriah (2009). *Subjective consciousness.* Oxford: Oxford.

Locke, John (1690/1975). *An essay concerning human understanding*. Oxford: Oxford.

McGeer, Victoria (1996). Is "self-knowledge" an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy, 93,* 483–515.

McGeer, Victoria (2008). The moral development of first-person authority. *European Journal of Philosophy, 16,* 81–108.

Moran, Richard (2001). *Authority and estrangement.* Princeton: Princeton.

Nichols, Shaun, and Stephen P. Stich (2003). *Mindreading*. Oxford: Oxford.

Noë, Alva (2004). *Action in perception*. Cambridge, MA: MIT.

Papineau, David (2002). *Thinking about consciousness*. Oxford: Oxford.

Pashler, Harold E. (1998). *The psychology of attention.* Cambridge, MA: MIT.

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 J. (2004a). The fractionation of introspection. *Journal of Consciousness Studies, 11* (no. 7–8), 40–57.

Prinz, Jesse J. (2004b). *Gut reactions.* New York: Oxford.

Prinz, Jesse J. (2007). All consciousness is perceptual. In *Contemporary debates in philosophy of mind,* Brian McLaughlin and Jonathan Cohen (eds.). Malden, MA: Blackwell.

Robinson, William S. (2005). Thoughts without distinctive non-imagistic phenomenology. *Philosophy and Phenomenological Research, 70,* 534–60.

Rosenthal, David M. (2005). *Consciousness and mind.* Oxford: Oxford.

Schachter, Stanley, and Jerome E. Singer (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69,* 379-399.

Scherer, Klaus R. (1984). On the nature and function of emotion: A component process approach. In *Approaches to emotion,* Klaus R. Scherer and Paul Ekman (eds). Hillsdale, NJ: Erlbaum.

Schwitzgebel, Eric (2002). How well do we know our own conscious experience? The case of visual imagery. *Journal of Consciousness Studies, 9* (no. 5–6), 35–53.

Schwitzgebel, Eric (2005). Difference tone training. *Psyche*, 11 (no. 6).

Schwitzgebel, Eric (2008). The unreliability of naive introspection. *Philosophical Review, 117,* 245–273.

Schwitzgebel, Eric (2010). Introspection. *Stanford Encyclopedia of Philosophy.*

Schwitzgebel, Eric (forthcoming). *Perplexities of consciousness.* Cambridge, MA: MIT.

Schwitzgebel, Eric (submitted). Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief.

Shoemaker, Sydney (1996). *The first-person perspective and other essays.* Cambridge: Cambridge.

Siewert, Charles (1998). *The significance of consciousness*. Princeton: Princeton.

Siewert, Charles (2004). Is experience transparent? *Philosophical Studies, 117,* 15–41.

Titchener, E.B. (1901–1905). *Experimental psychology*, New York: Macmillan.

Titchener, E.G. (1912). The schema of introspection. *American Journal of Psychology, 23,* 485-508.

Tye, Michael (2000). *Consciousness, color, and content*. Cambridge, MA: MIT.

Varela, Francisco J. (1999). The specious present: A neurophenomenology of time consciousness. In *Naturalizing phenomenology,* Jean Petitot et al. (eds.). Stanford CA: Stanford.

Wittgenstein, Ludwig (1953/1968). *Philosophical investigations*, 3rd ed. G.E.M. Anscombe (trans.). New York: Macmillan.