On the Unreliability of Introspection

For a *Philosophical Studies* book symposium on Eric Schwitzgebel's,

Perplexities of Consciousness.

1. The Unreliability Thesis

In his provocative and engaging new book, *Perplexities of Consciousness*, Eric Schwitzgebel argues that introspection is unreliable in the sense that we are prone to ignorance and error in making introspective judgments about our own conscious experience. Let us call this *the unreliability thesis*.

Schwitzgebel's argument for the unreliability thesis does not rely on abstract, theoretical principles, but instead proceeds through a vivid discussion of an intriguing series of case studies, including dreams, depth-perception, echolocation, imagery, inattention, emotion, conscious thought and phosphenes. In each case, the aim is to persuade us that we encounter either disagreement or uncertainty about how to describe our own conscious experience. This serves as the basis for a more general conclusion that our best introspective judgments about the most general features of our conscious experience are unreliable in the sense that they are vulnerable to ignorance and error.

My aim here is not to challenge Schwitzgebel's argument for the unreliability thesis by disputing his treatment of individual cases or by questioning the extrapolation from unreliability in these cases to unreliability across the board.[1] Instead, I propose to concede the unreliability thesis, at least for the sake of the argument, in order to consider its broader philosophical implications.

Schwitzgebel makes two distinct claims about the philosophical significance of the unreliability thesis. In chapter six, he argues that the unreliability of introspection has important methodological consequences for the empirical science of consciousness. Introspective reports play an indispensable role as a source of evidence for scientific theories of the neural basis of consciousness. So, if these introspective reports are sufficiently unreliable, Schwitzgebel concludes, "a methodologically well justified scientific consensus on a theory of consciousness may be beyond our reach" (2011: 115).

In chapter seven, Schwitzgebel makes a further suggestion that the unreliability thesis has important epistemological consequences for the prospects of a broadly Cartesian epistemology on which introspective judgments about our own conscious experience play a foundational epistemic role. Thus, he writes:

> Descartes thought, or is often portrayed as thinking, that we know our own experience first and most directly and then infer from it to the external world. If that is right – if our judgments about the outside world, to be trustworthy, must be grounded in sound judgments about our experiences – then our epistemic situation is dire indeed. However, I see no reason to accept any such introspective foundationalism. Indeed, I suspect that the opposite is nearer the truth: Our judgments about the world tend to drive our judgments about our experience. Properly so, since the former are the more secure. (2011: 137)

In these comments, I will be primarily concerned with Schwitzgebel's claims about the epistemological significance of the unreliability thesis.[2] To anticipate, I will be arguing that the unreliability thesis can be reconciled with a broadly Cartesian epistemology and hence that it does not have the revolutionary epistemological consequences that Schwitzgebel claims it does.

First, though, I want to raise a question about how radical the unreliability thesis is intended to be. On the one hand, it is clear that Schwitzgebel rejects *infallibilism* about introspection – that is, the thesis that introspection is *always* a source of knowledge. However, despite its historical association with Descartes, infallibilism has few, if any, contemporary proponents. On the other hand, it is equally clear that Schwitzgebel rejects *skepticism* about introspection – that is, the thesis that introspection is *never* a source of knowledge; as he admits, "I am not an utter skeptic" (2011: 139). Schwitzgebel rejects these two extreme views in favour of the more moderate view that introspection is sometimes, but not always, a source of knowledge. On this view, introspection is neither perfectly reliable nor perfectly unreliable. All this seems true, but not particularly controversial. Echoing Schwitzgebel (2011: 119), one might be tempted to ask, "Where are the firebrands?"

It strikes me that what is most radical in Schwitzgebel's work on introspection is his opposition to what we might call the *epistemic privilege thesis* – that is, the thesis that introspection is epistemically privileged in the sense that it is more reliable, or less vulnerable to ignorance and error, than other ways of knowing about the world, such as sensory perception. In one form or another, the epistemic privilege thesis is still widely endorsed and continues to exert a powerful influence

in contemporary epistemology. In light of this, Schwitzgebel's unreliability thesis is perhaps best stated in comparative terms: introspection is *no more reliable* than other ways of knowing about the world, including sensory perception; indeed, if anything, it is *less reliable*. This comparative claim is what motivates Schwitzgebel's rejection of Cartesian epistemology, which he sums up as follows: "Descartes, I think, had it quite backward when he said the mind – including especially current conscious experience – was better known than the outside world" (2011: 136).

Schwitzgebel's arguments for the unreliability thesis constitute a powerful challenge to contemporary incarnations of Cartesian epistemology that rely on some version of the epistemic privilege thesis. My aim here is to respond to this challenge by arguing that the epistemic privilege thesis can be understood in a way that is consistent with Schwitzgebel's unreliability thesis. If so, then the unreliability thesis need not have the revolutionary consequences that Schwitzgebel claims it does for the prospects of a broadly Cartesian epistemology.

2. Cartesian Skepticism

Cartesian skepticism provides a useful starting point for thinking about the epistemological differences between introspection and other ways of knowing about the world, such as sensory perception. After all, as Schwitzgebel observes, "Current conscious experience is generally the last refuge of the skeptic against uncertainty" (2011: 117). Descartes' skeptical hypothesis that I am dreaming, or that I am deceived by an evil demon, is incompatible with my perceptual knowledge of the external world, but not with my introspective knowledge of my own conscious

experience. As Descartes puts the point in the Second Meditation, "Because I may be dreaming, I can't say for sure that I now see the flames, hear the wood crackling, and feel the heat of the fire; but I certainly *seem* to see, to hear, and to be warmed."

Nevertheless, Schwitzgebel claims that we can construct skeptical scenarios that threaten our introspective knowledge of our own conscious experience in much the same way as the Cartesian skeptical scenarios threaten our perceptual knowledge of the external world. Thus, he writes:

> If you admit the possibility that you are dreaming, I think you should admit the possibility that your judgment that you are having red phenomenology is a piece of delirium not accompanied by any reddish phenomenology. (2011: 124)

Schwitzgebel is certainly correct that such cases of delirium are possible; indeed, an actual case is found in patients with Anton's syndrome, who are blind, but believe that they can see.[3] And yet, as I explain below, these cases are quite unlike Cartesian skeptical scenarios in ways that reflect an important epistemic asymmetry between perception and introspection.

In the First Meditation, Descartes invites us to consider the hypothesis that we are like "brain-damaged madmen who are convinced they are kings when really they are paupers, or say they are dressed in purple when they are naked, or that they are pumpkins, or made of glass." It is no accident that Descartes does not rest here, since the hypothesis that I am a brain-damaged madman functions quite

differently in skeptical arguments from the more canonical form of Cartesian skeptical hypothesis that I am dreaming or deceived by an evil demon. The crucial difference is that a madman's beliefs, unlike my own, are not justified by the evidence of the senses. If I am dreaming or deceived by an evil demon, then my beliefs are false, or otherwise fail to be knowledge, although they are properly based on the justifying evidence of the senses. In the madman scenario, by contrast, my beliefs are false, or otherwise fail to be knowledge, because they are *not* properly based on the justifying evidence of the senses.

In order to mark this contrast, we need to introduce a distinction between *brute errors*, which are justified false beliefs that are properly based on justifying evidence, and *basing errors*, which are unjustified false beliefs that are not properly based on justifying evidence.[4] In this terminology, Descartes' dreaming and evil demon scenarios are cases involving brute errors, whereas the madman scenario merely involves basing errors. Similarly, the introspective errors involved in Schwitzgebel's delirium scenario and in Anton's syndrome are basing errors, rather than brute errors. In all of these cases, the effect of mental illness is that its victims form false beliefs in a way that is not properly based on the justifying evidence that is provided by their own experience.

In order to extend the threat of Cartesian skepticism from our perceptual knowledge of the external world to our introspective knowledge of our own conscious experience, we need skeptical scenarios for the case of introspection that involve brute errors, as opposed to basing errors. However, Schwitzgebel does not succeed in providing skeptical scenarios of this kind. Moreover, in my view, this is

no accident. In what follows, I will argue for a version of the epistemic privilege thesis on which introspection, unlike perception, is immune from the possibility of brute ignorance and brute error. I claim that all of Schwitzgebel's examples of the unreliability of introspection can be explained away as basing errors, rather than brute errors. If that is right, then Schwitzgebel's unreliability thesis can be reconciled with a version of the Cartesian idea that there is a fundamental epistemic asymmetry between perception and introspection.

3. The Simple Theory of Introspection

Perception is subject to brute errors in which one forms justified, but false beliefs about the world on the basis of perceptual experience. After all, perceptual experience is *representational*: in veridical cases, it represents the way the world is, and in cases of illusion and hallucination, it *misrepresents* the way the world is. Moreover, in all these cases, one's perceptual experience provides defeasible justification to believe that the world is the way that it is represented to be. So, in cases of perceptual misrepresentation, one's perceptual experience provides defeasible justification to believe false propositions about the way the world is.

Introspection, by contrast, is immune from brute errors in which one forms justified, but false beliefs about one's experience on the basis of introspective representations that misrepresent one's experience. On some views, this is because introspective beliefs about one's experience are formed on the basis of introspective representations that are incapable of misrepresentation. For instance, it is sometimes claimed that conscious experience is *self-representing* in the sense that

all conscious experiences essentially represent themselves.[5] In my view, by contrast, introspection is not a matter of forming beliefs on the basis of representations of conscious experience at all. According to *quasi-perceptual theories* of introspection, one's introspective beliefs about one's experiences are caused and justified by representations of those experiences. Instead, I propose a *simple theory* of introspection on which one's introspective beliefs about one's own experiences are caused and justified by the experiences they are about.[6]

According to the simple theory, introspection is a distinctive way of knowing about one's experience that one has just by virtue of having that experience. Similarly, introspective justification is a distinctive kind of justification that one has to form beliefs about one's experience just by virtue of having that experience. There is no further requirement that one must *represent* one's experience in order to have introspective justification to believe that one has that experience. On the simple theory, the source of one's introspective justification to believe that one has a certain kind of experience is constituted by that experience itself, rather than any representation of that experience.

A consequence of the simple theory is that there cannot be brute errors in the case of introspection because the source of introspective justification is identical with its subject matter. Since one's introspective justification to believe that one has a certain experience has its source in the fact that one has that experience, it follows that one cannot have introspective justification to believe false propositions about one's experience. Thus, if the simple theory is true, then all experiences satisfy the following epistemic privilege thesis:

> The epistemic privilege thesis: one has introspective justification to believe that one has an experience E if and only if one has E.

However, the epistemic privilege thesis is distinct from, and does not entail, the doxastic privilege thesis below:

> The doxastic privilege thesis: One has an introspectively justified belief that one has an experience E if and only if one has E.

After all, one may have introspective justification to believe that one has a certain experience even if one does not use it, and perhaps cannot use it, in forming an introspectively justified belief. Since we are non-ideal agents, our introspective beliefs are not always formed in a way that is perfectly sensitive to the experiences that provide us with introspective justification. Hence, there can be introspective basing errors in which one falsely believes that one has an experience of a certain kind on the basis of an experience of a different kind, although there can be no brute errors, since beliefs formed in this way are not introspectively justified.

To illustrate the point, consider the initiation case in which you are threatened with a red hot poker and then touched on the neck with an ice cube so that you are tricked into mistaking the mildly unpleasant sensation of cold for an intensely painful sensation of heat. This is an example of a basing error, since you have introspective justification to believe that your neck feels cold, although you are

tricked into forming a false and unjustified belief that your neck feels hot. Of course, we don't blame you for believing this, but the mere absence of blameworthiness is not sufficient for the presence of justification. We might even say that your belief is reasonable or justified by ordinary standards that take into consideration the extenuating circumstances and one's more general psychological limitations. By ideal standards, however, the most reasonable or justified belief to hold in the circumstances is that one's neck feels cold, since this is what one has introspective justification to believe in virtue of having the sensation in question.

In his discussion of various case studies, Schwitzgebel makes a compelling case that we are prone to widespread ignorance and error in forming introspective judgments about our own conscious experience. This is enough to undermine the doxastic privilege thesis, but it leaves the epistemic privilege thesis intact, so long as Schwitzgebel's examples of ignorance and error can be explained away as failures of basing in which one fails to believe what one has introspective justification to believe. For instance, in chapter six, Schwitzgebel considers the question of whether we have experience in the absence of attention, such as a constant tactile experience of our feet in our shoes. On this question, opinion is divided: some say we do, others say we don't, while yet more remain undecided. This is evidence enough of widespread ignorance and error in the domain of introspection, but can we plausibly maintain that this ignorance and error is unjustified at least by ideal standards, if not by ordinary standards?

It might seem to be an implausible prediction of the simple theory that in hard debates about the nature of conscious experience, one side is always more

justified than the other, namely the side that happens to get the right answer. It is more plausible that there is approximate parity in respect of justification between the opposing parties in these debates. Indeed, given this parity, it might seem that the most reasonable or justified response to the disagreement would be to withhold belief altogether, rather than taking sides. According to the simple theory, however, if we have constant tactile experience of our feet in our shoes, then we have justification to believe that we do, and if we don't, we don't. So, it may be objected, the simple theory yields implausible predictions.[7]

In response, however, we need to draw two distinctions. The first distinction is between *propositional* and *doxastic* justification – that is, between which propositions one has justification to believe and which justified beliefs one has. A justified belief is one that is held in a way that is appropriately based on, or sensitive to, the presence of one's justification to hold the belief in question. The simple theory predicts an asymmetry in propositional justification, but not doxastic justification: in other words, we may have introspective justification to resolve a debate in one way or another, but we may be unable to exploit this in forming introspectively justified beliefs, since our doxastic dispositions may be insufficiently sensitive to the facts about conscious experience that determine which propositions we have introspective justification to believe.[8]

The second distinction is between justification by *ideal* standards and justification by *ordinary* standards. It may be that the response that is most justified by ideal standards is to resolve the debate in one direction or another, but if our contingent doxastic limitations prevent us from resolving the debate in a way that is

justified, then withholding belief may be the response that is most justified by ordinary standards.[9] Ordinary standards of justification take into account these contingent doxastic limitations, whereas ideal standards of justification abstract away from them.

In this way, the simple theory of introspection can be reconciled with Schwitzgebel's pessimism about its impotence in settling hard questions about the nature of conscious experience. We can explain the unreliability of our introspective beliefs in terms of our failure to use infallible introspective justification, rather than our success in using fallible introspective justification. By contrast, we cannot explain the unreliability of our perceptual beliefs in the same way, since it is so implausible to deny that perceptual illusions and hallucinations provide fallible justification by misrepresenting the environment. The simple theory explains this epistemic asymmetry between perception and introspection by appealing to the fact that perception, unlike introspection, is representational.

4. Conclusions and Further Discussion

Schwitzgebel makes a compelling case that introspection is unreliable in the sense that we are prone to ignorance and error in making introspective judgments about our own conscious experience. Nevertheless, I have argued that Schwitzgebel's unreliability thesis can be reconciled with a qualified version of the epistemic privilege thesis, according to which introspection is more reliable, and so less vulnerable to ignorance and error, than other ways of knowing about the world, such as sensory perception. In particular, it is consistent with the claim that

introspection, unlike perception, is immune from brute ignorance and error, although like perception, it is subject to ignorance and error that results from failures of basing. Therefore, I conclude that Schwitzgebel's unreliability thesis is less damaging for the prospects of Cartesian epistemology than he claims.

Before closing, I want to discuss three objections to this proposed reconciliation between Cartesian epistemology and Schwitzgebel's unreliability thesis. I should note that these objections raise large-scale issues that cannot be settled quickly. Nevertheless, I hope that my brief discussion will highlight some of the outstanding challenges that are raised by Schwitzgebel's work.

The first objection is that the epistemic privilege thesis depends upon the existence of a sharp distinction between introspection and other ways of knowing about the world, such as sensory perception. In response to this, Schwitzgebel (2011: 136-7; see also 2012) argues that we cannot disentangle the mechanisms and processes that underpin our introspective and perceptual beliefs. But if introspection and perception are not distinct kinds of psychological processes, then how can introspection be epistemically privileged with respect to perception?

This objection relies on the assumption that questions about the epistemology of introspection are hostage to the psychology of introspection. However, the simple theory is a theory of the nature of introspective justification, which makes introspective knowledge possible. It is not an account of the psychological mechanisms or processes that we use in forming introspectively justified beliefs and acquiring introspective knowledge. As far as the simple theory

is concerned, it is an open question how these beliefs are formed, so long as they are more or less directly sensitive to the experiences they are about.[10]

Even if we cannot draw a sharp distinction between perception and introspection at the level of psychological mechanisms, we can draw a sharp distinction at the level of epistemology. After all, the source of introspective justification is constituted by its subject matter, whereas the source of perceptual justification is constituted by representations of its subject matter. This is a crucial epistemological difference between perception and introspection which remains however much overlap there is between the psychological mechanisms and processes that underpin the formation of beliefs in perception and introspection.

The second objection is that, assuming a reliabilist theory of justification, Schwitzgebel's unreliability thesis is inconsistent with the epistemic privilege thesis defended here. According to reliabilism, which propositions one has justification to believe depends upon the reliability of one's doxastic dispositions. In particular, one has introspective justification to believe that one has an experience of a certain kind if and only if one has a reliable introspective mechanism that disposes one to believe that one has experiences of that kind. So, if one's introspective mechanisms are unreliable, then there will be cases of brute ignorance and brute error in which the experiences that one actually has come apart from the experiences that one has introspective justification to believe one has.

Of course, reliabilism is very controversial and it would be question-begging to assume it in the context of a critique of Cartesian epistemology. After all, Cartesian skeptical scenarios provide the inspiration for well known

counterexamples to reliabilism, such as the "new evil demon" scenario, in which mental duplicates are alike in which propositions they have justification to believe, although they differ in the reliability of their doxastic dispositions. These examples motivate a rejection of reliabilism in favour of mentalism, according to which one's mental states determine which propositions one has justification to believe in a way that does not depend on the reliability of their connections to the external world.

Assuming reliabilism, there is no obvious reason to accept the existence of an epistemic asymmetry between introspection and perception, since ignorance and error is equally possible in each case. Assuming mentalism, by contrast, we can draw a distinction between brute ignorance and error on the one hand and ignorance and error that results from failures of basing on the other. Given this distinction, there is no easy route from the unreliability thesis to a rejection of the epistemic asymmetry between perception and introspection. On the contrary, it is plausible that the mental differences between perception and introspection ground corresponding epistemological differences.

The third and final objection is that if introspective justification is defeasible, then it can be defeated by evidence that introspection is unreliable. In this way, Schwitzgebel's argument for the unreliability thesis can provide the basis of a more general argument for skepticism about introspection. The upshot is that introspection is no less vulnerable to the threat of skepticism than perception.

My inclination in response would be to deny the assumption that introspective justification is defeasible by evidence about its reliability. On this view, higher-order evidence about the reliability of introspection does not defeat one's

introspective justification for beliefs about one's own experience, but rather poses an obstacle that prevents non-ideal agents from exploiting their introspective justification in forming introspectively justified beliefs.[11] But since I do not have the space to defend that response here, let me simply note that introspective justification is not defeasible in the same way as perceptual justification. Since perception is representational, perceptual justification is defeasible by considerations that justify believing that one's perceptual experience misrepresents the environment. Introspection, by contrast, is not representational and so the justification that is provided by one's experience for believing that one has that experience cannot be defeated in the same way. So, even if introspection is vulnerable to skeptical arguments, it is not vulnerable to the standard forms of skeptical argument that apply in the case of perception.

In conclusion, Schwitzgebel's unreliability thesis has important implications for a wide range of issues in epistemology, but it would be premature to conclude that it constitutes a decisive case against Cartesian epistemology. As I have been arguing, the simple theory of introspection provides the resources for reconciling Schwitzgebel's unreliability thesis with a broadly Cartesian epistemology.[12]

References

Bayne, T. and Spener, M. 2010. Introspective Humility. *Philosophical Issues* 20: 1-22.

Burge, T. 1988. Individualism and Self-Knowledge. *Journal of Philosophy* 85: 649-63.

Kriegel, U. and Horgan, T. 2007. Phenomenal Epistemology: What is Consciousness That We May Know It So Well? Philosophical Issues 17: 123-144.

MacPherson, F. 2010. A Disjunctive Theory of Introspection: A Reflection on Zombies and Anton's Syndrome. *Philosophical Issues* 20: 226-65.

Schwitzgebel, E. 2011. *Perplexities of Consciousness*, Oxford University Press.

Schwitzgebel, E. 2012. Introspection, What? In *Introspection and Consciousness*, eds. D. Smithies & D. Stoljar, Oxford University Press.

Smithies, D. 2012a. A Simple Theory of Introspection. In *Introspection and Consciousness*, eds. D. Smithies & D. Stoljar, Oxford University Press.

Smithies, D. 2012b. Mentalism and Epistemic Transparency. *The Australasian Journal of Philosophy* 90: 723-42.

Smithies, D. Forthcoming. Epistemic Idealization and Higher-Order Evidence. *Synthese*.

---

[1] See Bayne and Spener (2010) and Spener (this volume) for defence of the claim that introspection is reliable within a restricted domain of operation.
[2] See Kriegel (this volume) for further discussion of the methodological implications of the unreliability thesis for the science of consciousness.
[3] See MacPherson (2010: 234-40) for a defence of the claim that patients with Anton's syndrome do not have visual experiences, but merely report falsely that they do.
[4] Burge (1988: 657) uses the term 'brute error' in a distinct, but related sense to denote errors that "do not result from any sort of carelessness, malfunction, or irrationality on our part."
[5] See Horgan and Kriegel (2007) for an exposition of this view.
[6] See Smithies (2012a) for an extended discussion and defence of the simple theory of introspection.
[7] I am grateful to Eric Schwitzgebel for pressing me to consider this objection.
[8] See Smithies (2012b) for a similar diagnosis of the problem of the speckled hen. If one's perceptual experience represents that a hen is 48-speckled, or more realistically, that it has a determinate shade, red-48, then this is what one has justification to believe. Nevertheless, one may be unable to form a justified belief that the hen is red-48, since one's doxastic dispositions are not sufficiently sensitive to the distinction between representing red-48 versus red-47 or red-49.
[9] Compare: it may be justified by ordinary standards to withhold belief in logical truths, but if probabilistic coherence is ideally justified, then it is justified by ideal standards to be certain of all logical truths.
[10] This is not something that Schwitzgebel disputes. See Schwitzgebel (2012: 42-3) for the claim that introspective judgments must reflect, or at least aim to reflect, "some relatively direct sensitivity to the target state".
[11] See Smithies (forthcoming) for this account of higher-order evidence.
[12] Thanks to Eric Schwitzgebel, Daniel Stoljar and an audience at the Pacific Meeting of the American Philosophical Association in April 2012 for helpful comments and discussion.