A Simple Theory of Introspection

Declan Smithies

1. What is Introspection?

What is introspection? Etymologically, 'introspection' means *looking within*. However, it cannot be assumed from the outset that introspection is best explained as a form of inner perception. A more neutral starting point is to use the term 'introspection' as a placeholder for the distinctive way (whatever it is) in which we know our own minds. This is to make the plausible assumption that there is a distinctively introspective way in which we know our own minds, but without begging important questions about the nature of this introspective way of knowing. Moreover, it defines the agenda for a theory of introspection – that is, to describe and explain what exactly is distinctive about the introspective way in which we know our own minds.

We can begin by noting that introspection has a distinctive *subject matter*. We will return in due course to the question of how exactly to demarcate its subject matter, but to a first approximation, introspection is a way of knowing about one's current mental states (events, processes, and so on), as opposed to the mental states of others, one's past mental states, or the states of one's body, such as one's height or weight. However, introspection cannot be defined solely in terms of its subject matter without collapsing the distinction between knowing about one's mental states by introspection and knowing about one's mental states in some other way. So, if introspection is a distinctive way of knowing about one's mental states, then we need to consider what sets it apart from other ways of knowing about the same subject matter.

Following Alex Byrne (2005), we can separate two dimensions along which introspection may be distinguished from other ways of knowing about the world. First, introspection is

peculiar in the sense that it is different in certain epistemological respects from other ways of knowing about its subject matter. And second, introspection is *privileged* in the sense that it is better in certain epistemological respects than other ways of knowing about its subject matter. A plausible constraint of adequacy on any theory of introspection is that it should describe and explain the sense in which introspection is both peculiar and privileged.

My primary goal in this paper is to defend what I call *the simple theory of introspection*. According to the simple theory, introspection is a distinctive way of knowing that one is in a certain mental state, which one has just by virtue of being in that mental state. The simple theory is motivated in part by reflection on examples. For instance, if I am in pain, then I have a distinctive way of knowing that I am in pain just by virtue of the fact that I am in pain. Similarly, if I am thinking about rhubarb, then I have a distinctive way of knowing that I am thinking about rhubarb just by virtue of the fact that I am thinking about rhubarb.[1]

Moreover, the simple theory explains what is distinctive about the introspective way in which we know about our own mental states. Introspection is peculiar in the sense that it is a way of knowing about one's own mental states that is not available to anyone else. If others know that I am in pain, then they know this by testimony or by making inferences from observation of my behaviour or the activity of my brain, whereas I am in a position to know that I am in pain just by virtue of being in pain. Moreover, introspection is privileged in the sense that it is a way of knowing about one's mental states that is always available in one's own case. Others are in a position to know that I am in pain only if they have sufficient evidence from observation, inference, or testimony, whereas I am always in a position to know that I am in pain just by virtue of being in pain. The simple theory therefore provides an elegant and plausible

explanation of the way in which introspection is epistemologically different from, and better than, other ways of knowing about the world.

Any theory will agree that one has an introspective way of knowing that one is in a certain mental state only if one is in that mental state, since knowledge is factive. However, many theories, including inner sense theories and reliability theories, impose further requirements to the effect that one has an introspective way of knowing that one is in a certain mental state only if one has an inner representation of that mental state or an introspective mechanism that reliably tracks that mental state. The simple theory, by contrast, claims that if one is in a certain mental state, then one has a way of knowing that one is in that mental state *just by virtue of being in that mental state*. However, the mere fact that *p* is not in general sufficient to provide one with a way of knowing that *p*. So, how is the mere fact that one is in a certain mental state sufficient to provide one with a way of knowing that one is in that mental state?

The simple theory explains the distinctive features of *introspective knowledge* in terms of *introspective justification*. Introspective justification, according to the simple theory, is the distinctive kind of justification that one has to believe that one is in a certain mental state, which one has just by virtue of being in that mental state.[2] As such, introspective justification is fundamentally different in character from other kinds of justification, including perceptual justification. For instance, one does not have perceptual justification to believe that the world is in a certain state just virtue of the fact that the world is in that state, but rather by virtue of one's perceptual experience, which *represents* that the world is in that state. The distinguishing feature of introspective justification is that *its source is identical with its subject matter*. Arguably, this has the further consequence that introspective justification, unlike perceptual justification, is (i) self-intimated; (ii) infallible; (iii) indefeasible; and (iv) immune from Gettier cases. If so, then

having introspective justification to believe that one is in a certain mental state is sufficient to provide one with a way of knowing that one is in that mental state. Let us briefly consider each of these features in turn.

First, introspective justification is *self-intimated* in the sense that if one is in a certain mental state, then one has introspective justification to believe that one is in that mental state. By contrast, perceptual justification is not self-intimated, since it is not the case that if the world is in a certain state, then one has perceptual justification to believe that it is in that state. Rather, one has perceptual justification to believe that the world is in a certain state only if one has a perceptual experience, which *represents* that the world is in that state. This has the interesting consequence that if ideal rationality involves believing every proposition that one has justification to believe, then ideal rationality is compatible with ignorance about the external world, but not with ignorance about one's own mental states.[3]

Second, introspective justification is *infallible* in the sense that if one has introspective justification to believe that one is in a certain mental state, then one is in that mental state. Again, perceptual justification is not infallible, since it is not the case that if one has perceptual justification to believe that the world is in a certain state, then it is in that state. After all, there are possible cases of perceptual illusion and hallucination in which one's perceptual experience *misrepresents* the world and so one has justification to believe falsely that the world is in a certain state. However, there are no possible cases of introspective illusion or hallucination, since the source of one's introspective justification is identical with its subject matter. As we shall see, there are possible cases in which one has false beliefs about one's own mental states – for instance, in the famous initiation trick, one is duped into mistaking a sensation of cold for a sensation of pain – but according to the simple theory, there are no such cases in which one has

false beliefs that are *introspectively justified*. As before, this implies that ideal rationality is compatible with error about the external world, but not with error about one's own mental states.

Third, introspective justification is *indefeasible* in the sense that it cannot be defeated by justification to believe anything else. Perceptual justification, on the other hand, can be defeated by justification to believe that one's perceptual experience is the result of illusion or hallucination. According to the simple theory, however, introspective justification cannot be defeated in this way, since it is a priori that there are no possible cases of introspective illusion or hallucination. Moreover, it is not clear that there is any other way in which introspective justification can be defeated. David Armstrong (1968: 109) gives the example of a brain technician, who informs me on the basis of neural evidence that I believe falsely that I seem to be seeing something green. But if my belief is false, then I have no introspective justification to be defeated; and if my belief is true, then it is more plausible to suppose that my introspective justification defeats the testimony of the brain technician, rather than vice versa.[4]

Fourth, introspective justification is immune from *Gettier cases* in the sense that if one has an introspectively justified belief that one is in a certain mental state, then one knows by introspection that one is in that mental state. Perceptual justification, of course, is subject to Gettier cases in which one has a true belief that is perceptually justified, but fails to constitute knowledge. The standard recipe for constructing Gettier cases is to begin with a case in which a justified belief is false owing to bad luck and then to add an element of good luck that counteracts the bad luck and thereby makes the justified belief true, but luckily so. However, this recipe cannot be applied unless justification is fallible, which explains why it applies in the case of perceptual justification, but not introspective justification.[5] Moreover, it is not clear that there is any other recipe for generating Gettier cases for introspectively justified beliefs.[6]

In summary, the simple theory claims that some mental states are *introspectively accessible* in the sense that being in a certain mental state is necessary and sufficient for having introspective justification to believe that one is in that mental state and thereby for having an introspective way of knowing that one is in that mental state. This claim is encapsulated in the following introspective accessibility thesis:

· **The introspective accessibility thesis:** for some mental states M, necessarily, one is in M if and only if one has introspective justification to believe that one is in M and one thereby has an introspective way of knowing that one is in M.

However, the introspective accessibility thesis raises further questions of its own.[7] First, there is *the generalization question*: which mental states are introspectively accessible in the relevant sense? And second, there is *the explanatory question*: what explains why some mental states are introspectively accessible, rather than others, or none at all?

The simple theory is motivated initially by reflecting on examples of introspective knowledge and the way in which it explains their peculiar and privileged status. However, opponents of the simple theory may deny that any mental states are introspectively accessible in the sense articulated by the simple theory. Therefore, a full defence of the simple theory requires answering the generalization question and the explanatory question in such a way as to provide a theoretical rationale for the claim that certain mental states are introspectively accessible, rather than others, or none at all. My strategy in this paper is to argue for the conditional claim that if access internalism is true, then the accessibility of facts about justification is best explained by the introspective accessibility of facts about one's mental states. Therefore, I conclude that access internalism provides a theoretical rationale for the simple theory.

The overall plan for the paper is as follows. In section two, I elaborate the simple theory by situating it within the context of a distinction between epistemic and doxastic versions of constitutivism. In sections three and four, I answer the generalization question by arguing that mental states are introspectively accessible if and only if they are phenomenally individuated. In sections five and six, I answer the explanatory question by arguing that mental states are introspectively accessible if and only if they are among the determinants of justification. Finally, in section seven, I conclude by reflecting on the nature of the relationship between rationality and introspective self-knowledge.

2. Epistemic and Doxastic Constitutivism

Constitutivism is the thesis that there is a constitutively necessary connection between certain mental states and our cognitive or epistemic access to those mental states. It is usually formulated in terms of an accessibility thesis of the following schematic form:

· **The schematic version:** for some mental states M, and some relation Φ, necessarily, one is in M if and only if one Φs that one is in M.

The accessibility thesis is a biconditional that conjoins a *self-intimation* thesis, which is the conditional that goes from left to right, and an *infallibility* thesis, which is the conditional that goes from right to left. The self-intimation rules out the possibility of certain kinds of ignorance, whereas the infallibility thesis rules out the possibility of certain kinds of error.

The schematic version of the accessibility thesis raises various questions. Which mental states, M, are accessible? Which are the modes of presentation under which are they accessible? And what is the cognitive or epistemic relation, Φ, in terms of which accessibility is defined?

According to doxastic versions of constitutivism, there is a constitutively necessary connection between certain mental states and our *beliefs* about those mental states. In its simplest incarnation, this is captured by the following doxastic version of the accessibility thesis:

- **The doxastic version:** for some mental states M, necessarily, one is in M if and only if one believes that one is in M.

Unfortunately, not even paradigm examples of phenomenally conscious mental states, such as feeling pain and feeling cold, are doxastically accessible in this sense. Even if we restrict the scope of the thesis to phenomenally conscious states, which are self-ascribed under first-person, present-tense phenomenal modes of presentation, it is easy to generate counterexamples.

First, there are cases of *conceptual poverty*, in which one is in a phenomenal state, but one does not believe that one is in that state because one does not possess the requisite concepts. For example, an infant feels pain, but it lacks the conceptual abilities required to believe that it feels pain. Second, there are cases of *inattention*, in which one is in a phenomenal state, but one does not believe that one is in that state because one's attention is distracted. For example, an athlete incurs a painful injury while playing football, but does not notice the pain until the game is over.[8] Third, there are cases of *misclassification*, in which one is in a phenomenal state, but one does not believe that one is in that state because one misclassifies it. For instance, a student in an initiation trick is threatened with a red hot poker and then touched with an ice-cube, so that he is duped into mistaking the mildly unpleasant feeling of cold for an intensely painful feeling of heat. In that case, he feels cold, but he does not believe that he feels cold; and moreover, he believes that he feels pain, but he does not in fact feel pain.[9]

How might we reformulate the accessibility thesis in such a way as to avoid these counterexamples? A common strategy is to retain a doxastic version of the accessibility thesis,

while imposing further restrictions designed to rule out the various counterexamples. My own strategy, by contrast, is to reformulate the accessibility thesis in epistemic, rather than doxastic, terms. I do not deny that there is any way of restricting the doxastic version of the accessibility thesis in such a way as to make it immune to counterexamples. However, I claim that any such restricted version of the doxastic accessibility thesis can be explained as a consequence of an unrestricted version of the epistemic accessibility thesis, which is more fundamental in the order of philosophical explanation. Or so I will argue in what follows.

The simple theory, as we have seen, entails an epistemic version of the accessibility thesis, according to which there is a constitutively necessary connection between certain mental states and our epistemic access to those mental states:

· **The epistemic version:** for some mental states M, necessarily, one is in M if and only if one has introspective justification to believe that one is in M and one thereby has an introspective way of knowing that one is in M.

The epistemic version of the accessibility thesis does not entail the doxastic version, since one may have justification to believe a proposition without using it to acquire justified belief; and likewise, one may have a way of knowing a proposition without using it in acquiring knowledge. Moreover, the epistemic version avoids the counterexamples to the doxastic version, since the limitations of one's doxastic capacities do not impose corresponding limits on which propositions one has justification to believe.[10] It may be that one has justification to believe a proposition even if one lacks the doxastic capacities required to use it in acquiring justified belief. Similarly, it may be that one has a way of knowing a proposition even if one lacks the doxastic capacities required to use it in acquiring knowledge. Psychological facts about one's concepts, attention, powers of discrimination, and other rational capacities, do not constrain

which propositions one has justification to believe, but only which of these justifications one has a capacity to use in forming justified beliefs. Similarly, these facts do not constrain which propositions one has a way of knowing, but only which of these ways of knowing one has a capacity to use in acquiring knowledge. On this view, one's doxastic access to one's mental states may be more limited than one's epistemic access to one's mental states in ways that reflect contingent facts about the limitations of one's doxastic capacities. This is because the epistemic version of the accessibility thesis articulates an epistemic ideal, which abstracts away from contingent facts about the limitations of one's doxastic capacities.

If there are no doxastic limits on which propositions one has justification to believe, then the epistemic version of the accessibility thesis needs no further restriction in order to avoid counterexamples involving conceptual poverty, inattention or misclassification.[11] As a result, it applies more generally than doxastic versions, which do need further restriction. Moreover, as I will argue in what follows, the epistemic version is more fundamental in the order of philosophical explanation, since it explains the plausibility of various different strategies for imposing restrictions on the doxastic version of the accessibility thesis.

One influential strategy is to abandon the self-intimation thesis altogether and to restrict the infallibility thesis to a class of demonstrative beliefs about one's phenomenal states. Arguably, some beliefs about phenomenal states are infallible because they self-ascribe a phenomenal state under a *demonstrative mode of presentation* that cannot be entertained unless one is in the phenomenal state in question.[12] This motivates the following restriction on the doxastic version of the infallibility thesis:

· **The restricted infallibility version:** for some mental states M, necessarily, if one believes that one is in M under a demonstrative mode of presentation, then one is in M.

However, this restricted version of the doxastic infallibility thesis can be explained as a consequence of the epistemic accessibility thesis. If there are some demonstrative propositions about one's phenomenal states that cannot be believed unless they are true, then it is equally plausible that they cannot be believed unless one has introspective justification to believe them. The assumption here is that the conditions that enable one to believe a demonstrative proposition of the relevant kind are also sufficient to provide one with introspective justification to believe it.[13] But if the epistemic accessibility thesis is true, then one has introspective justification to believe a proposition only if it is true. Therefore, the infallibility of demonstrative beliefs about one's phenomenal states may be explained as a consequence of the infallibility of introspective justification together with the existence of epistemic constraints on demonstrative concepts and demonstrative beliefs about one's phenomenal states.

Another influential strategy is to avoid counterexamples by embedding the doxastic version of the accessibility thesis within the scope of a conditional, which restricts its application to subjects that have the requisite concepts, attention, and rationality:[14]

- **The conditional version:** for some mental states M, necessarily, if one is conceptually competent, attentive, and rational, then one is in M if and only if one believes that one is in M.

However, these restrictions are not sufficient to avoid cases of misclassification unless we assume a highly idealized conception of rationality, according to which the rational subject believes that she is in a certain mental state if and only she has introspective justification to believe that she is in that mental state. For instance, the victim of the initiation trick may be rational by ordinary standards, but he is not ideally rational, since he does not believe what he has introspective justification to believe – namely, that she feels cold, rather than pain. But if the

epistemic accessibility thesis is true, then one has introspective justification to believe that one is in a certain mental state if and only if one is in that mental state. Therefore, it follows that if one is ideally rational, then one believes that one is in a mental state if and only if one is in that mental state. Hence, the conditional version of the doxastic accessibility thesis can be explained as a consequence of the epistemic accessibility thesis.

Another reaction is to interpret the conditional version of the doxastic accessibility thesis in such a way that it allows for exceptions.[15] If we assume an ordinary conception of normal human rationality, rather than a more highly idealized conception, then it may be that rationality imposes limits on the possibility of ignorance and error about one's own mental states without requiring absolute omniscience or infallibility. Once again, however, this weakened version of the doxastic accessibility thesis can be explained by appeal to the accessibility thesis. Arguably, one's concepts of one's own mental states are epistemically individuated in the sense that one possesses those concepts if and only if one is disposed to use them in forming introspectively justified beliefs about one's mental states.[16] However, these possession conditions do not require achieving the rational ideal of omniscience and infallibility about one's own mental states, but merely some degree of approximation towards the rational ideal. Therefore, the limits on the possibility of ignorance and error about one's mental states can be explained by the self-intimation and infallibility of introspective justification together with the existence of epistemic constraints on possessing concepts of one's mental states.

Epistemic and doxastic versions of constitutivism stand in need of different kinds of explanation. Doxastic versions of constitutivism demand explanation in terms of psychological claims about the nature of our mental states, the nature of our beliefs about those mental states, and the nature of concepts, attention, and normal human rationality. For instance, doxastic

versions of constitutivism are sometimes explained by appeal to part-whole relations of constitution that hold between introspective beliefs and the mental states they are about.[17] Epistemic versions of constitutivism, on the other hand, should be explained in terms of epistemological, rather than psychological, claims. In particular, the epistemic version of the accessibility thesis is best explained by the simple theory of introspective knowledge and introspective justification.

The simple theory is a theory of the epistemology of introspection, rather than the psychology of introspection. It is an account 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 taking advantage of introspective justification and thereby acquiring introspective knowledge. As such, it makes no commitments about the reliability or unreliability of our beliefs about our own mental states. As far as the simple theory is concerned, it is an open question to what extent and by what mechanisms or processes we take advantage of what we have introspective justification to believe.[18]

As we have seen, however, the simple theory raises questions of its own. A full defence of the simple theory requires answering the generalization question and the explanatory question: which mental states are introspectively accessible and what explains why those mental states are introspectively accessible, rather than others, or none at all? My aim in the remaining sections of this paper is to provide answers to these questions.

3. The Generalization Question

Not all mental states are introspectively accessible. One's mental states, by any non-controversial criterion, include "subdoxastic" mental representations, which play a role in computational

explanations in cognitive science – for instance, Noam Chomsky's (1957) tacit knowledge of syntactic rules and David Marr's (1982) primal and 2.5D sketches.[19] Clearly, however, our justification to believe in the existence of these subdoxastic mental representations derives from scientific theory, rather than introspection. Therefore, we need some restriction on which mental states are introspectively accessible.

An initially tempting criterion is that a mental state is introspectively accessible if and only if it is *phenomenally conscious*. However, this criterion is too restrictive, since it excludes not only subdoxastic states, but also beliefs. Beliefs are standing states, which persist through time without making any ongoing contribution to one's phenomenology. For instance, my belief that Canberra is the capital of Australia persists whether or not I am consciously considering the matter and so does my second-order belief that I *believe* that Canberra is the capital of Australia. If these beliefs are justified, as we may assume, then we can ask what makes them justified at any given time. Crucially, however, there may be nothing in my stream of phenomenal consciousness that makes them justified at any given time. In some cases, my beliefs may be inferentially justified by other standing beliefs, but this is not always plausible, since I often forget the evidence on which my beliefs were originally formed.[20] In that case, the only plausible candidate for what makes my second-order belief justified is the mere presence of my first-order belief. In other words, I have introspective justification to believe that I believe that Canberra is the capital of Australia just by virtue of believing it.[21]

Any plausible answer to the generalization question must therefore be permissive enough to include beliefs, but also restrictive enough to exclude subdoxastic states. What we need, then, is a criterion that explains what beliefs, unlike subdoxastic states, have in common with phenomenally conscious states in virtue of which they are introspectively accessible. Richard

Rorty (1979, 22) expresses pessimism about the prospects for any unified account of beliefs and phenomenally conscious states when he writes, "The attempt to hitch pains and beliefs together seems ad hoc – they don't seem to have anything in common except our refusal to call them 'physical'." In what follows, however, I will suggest that this pessimism is unfounded.

One strategy for answering the generalization question is to propose some broadly functionalist criterion, according to which a mental state is introspectively accessible if and only if it plays a certain functional role. The challenge for proponents of this strategy is to identify some functional similarity that holds between beliefs and phenomenally conscious states, but not between beliefs and subdoxastic states. For instance, one might argue that beliefs are like phenomenally conscious states and unlike subdoxastic states insofar as they satisfy Ned's Block's functional criteria for access consciousness:

> A state is access conscious if it is poised for direct control of thought and action. To add more detail, a representation is A-conscious if it is poised for free use in reasoning and for direct "rational" control of action and speech. (1997, 382)

Since beliefs are typically access conscious, but not phenomenally conscious, one might answer the generalization question by replacing the appeal to phenomenal consciousness with access consciousness. On this proposal, a mental state is introspectively accessible if and only if it is *access conscious*.[22]

Counterexamples emerge when we consider hypothetical examples of subdoxastic states, which are access conscious, but not phenomenally conscious. In actual cases of blindsight, visual information is neither phenomenally conscious nor access conscious, since it is not poised for the direct control of speech and action except in forced choice conditions. In Block's (1997, 385-6) hypothetical case of *super-blindsight*, by contrast, visual information is access conscious, since it is poised for spontaneous use in the control of action and speech, although it is not phenomenally

conscious. Moreover, it is not introspectively accessible. Intuitively, the super-blindsighter does not have introspective justification to form beliefs about what is represented in her visual system any more than the regular blindsighter does. At best, she has justification to make inferences about what is represented in her visual system from observational data about her own spontaneous verbal and non-verbal behaviour. Therefore, access consciousness is not a sufficient condition for introspective accessibility.

One reaction is to impose a more demanding functional criterion, according to which a mental state is introspectively accessible if and only if it is not merely access conscious, but also *reflexively conscious*. A mental state M is reflexively conscious if and only if the subject has formed (or is disposed to form) a higher-order representation that she is in M, without relying on observation, inference, or testimony. According to higher-order theories of phenomenal consciousness, a mental state is phenomenally conscious if and only if it is reflexively conscious.[23] This gives rise to a dilemma for the current proposal. If higher-order theories are true, then reflexive consciousness is not a necessary condition for introspective accessibility, since beliefs are introspectively accessible, but not reflexively conscious. If higher-order theories are false, on the other hand, then reflexive consciousness is not a sufficient condition for introspective accessibility, since there are hypothetical examples of subdoxastic states that are reflexively conscious, but not introspectively accessible.

Consider the case of *hyper-blindsight*, which is just like super-blindsight, except that the subject has a reliable mechanism that is disposed to generate higher-order beliefs about phenomenally unconscious visual states without reliance on observation, inference, or testimony. Intuitively, the hyper-blindsighter does not have introspective justification to form beliefs about what is represented in her visual system any more than the super-blindsighter or the regular

blindsighter does. Certainly, she has a reliable disposition to form true beliefs about what is represented in her visual system, but this is not sufficient to make her beliefs introspectively justified. By analogy, the super-blindsighter has a reliable disposition to form true beliefs about stimuli in the blind hemi-field, but this is not sufficient to make them perceptually justified.[24] So why should we suppose that the hyper-blindsighter's non-inferential beliefs about what is represented in her visual system are any more justified than the super-blindsighter's non-inferential beliefs about what is located in her blind hemi-field?

In my view, there is no functional criterion for introspective accessibility that captures the distinction between beliefs and subdoxastic states except one that explicitly invokes relations to phenomenology. Beliefs are not phenomenally conscious states, but they are disposed to cause phenomenally conscious states of judgement. Moreover, I claim that beliefs are individuated wholly by their disposition to cause phenomenally conscious states of judgement. Subdoxastic states may also be disposed to cause phenomenally conscious states, but they are not individuated wholly by such dispositions. On the contrary, they are individuated in part by their disposition to play a role in computational processes which occur below the level of phenomenal consciousness. Let us say that a mental state is *phenomenally individuated* if and only if it is wholly individuated by its relations to phenomenally conscious states.[25] My proposal is that a mental state is introspectively accessible if and only if it is phenomenally individuated. In other words, my proposal is that phenomenal individuation is the criterion that explains why beliefs are introspectively accessible, whereas subdoxastic states are not.

Subdoxastic states are not individuated wholly by dispositions to cause phenomenally conscious states. Consider Martin Davies' (1989) hypothetical example in which subdoxastic states that embody tacit knowledge of syntax are disposed to cause phenomenally conscious

itches or tickles. These states are not individuated by their disposition to cause itches and tickles, but rather by their roles in syntactic processing. Similarly, mental representations in the computational theory of vision are not individuated solely by their disposition to cause phenomenally conscious visual experiences, but rather by their roles in visual processing. In principle, there could be subjects with exactly the same visual experiences, which are produced by different computational processes defined over different sets of subdoxastic visual states. Similarly, there could be subjects with exactly the same linguistic performance, and the same conscious states of linguistic understanding, but which are produced by different kinds of semantic and syntactic processing.[26]

Beliefs, on the other hand, are individuated wholly by their disposition to cause phenomenally conscious states of judgement. Judgement, as I use the term, is a phenomenally conscious state: there is something it is like to judge that a proposition is true. Moreover, the phenomenology of judgement is, in David Pitt's (2004) terminology, proprietary, distinctive and individuative.[27] It is *proprietary* in the sense what it's like to judge a proposition is different from what it's like to adopt a different attitude towards the same proposition; in other words, its phenomenology is attitude-specific. It is *distinctive* in the sense that what it's like to judge a proposition is different from what it's like to adopt the same attitude towards a different proposition; in other words, its phenomenology is content-specific. Moreover, it is *individuative* in the sense that judgement is individuated wholly by its content-specific and attitude-specific phenomenology.[28]

Beliefs are individuated wholly by their disposition to cause the proprietary, distinctive and individuative phenomenology of judgement. A mental state counts as a belief by virtue of its disposition to cause phenomenally conscious states with the proprietary phenomenology that is

18

specific to judgement. Moreover, a mental state counts as a belief in some specific propositional content by virtue of its disposition to cause phenomenally conscious states with the distinctive phenomenology that is specific to judging that content. The essence of belief is exhausted by its disposition to cause one to judge its propositional content: in other words, one believes a proposition if and only if (and in virtue of the fact that) one is disposed to judge the proposition in question.[29]

If belief is a disposition to judge, then not all judgements express beliefs and not all beliefs express themselves in judgement. A judgement does not express belief unless it manifests a stable disposition and a belief does not express itself in judgement unless the stable disposition is manifested. Nevertheless, it may be objected that a stable disposition to judge a proposition is neither necessary nor sufficient for believing it, since one's dispositions towards judgement sometimes fail to reflect what one really believes. Christopher Peacocke (1998, 90) gives the example of an academic on a hiring committee who is disposed to judge on good evidence that foreign degrees are equal in standard to domestic degrees, although her votes in hiring decisions reveal that she does not really believe this; indeed, what she really believes is that foreign degrees are inferior. In my view, however, Peacocke's example is misdescribed: the academic does not really believe that foreign degrees are inferior, although she acts *as if* she believes this. What motivates this redescription is that her irrationality is practical, not epistemic: it stems from the subject's actions, rather than her beliefs. After all, she knows full well that foreign degrees are equal in standard to domestic degrees: we may assume that she has studied the evidence carefully and formed her beliefs in a way that is appropriately responsive to the evidence. If so, then her rational failing is not that she has unjustified beliefs, but that her justified beliefs fail to exert an appropriate influence on her actions, which are consequently unjustified.

This is not to deny that Peacocke's academic has some attitude towards the proposition that foreign degrees are inferior to domestic degrees, which is functionally individuated by its role in disposing her to treat them as such. Rather, it is to deny that such an attitude is correctly described as 'belief'. And the substance behind this terminological point is that in such cases in which one's dispositions to judge a proposition come apart from one's dispositions to act as if it is true, our epistemic evaluations are sensitive to the former, but not the latter. [30]

In fact, for current purposes, the key point can be made without relying on this terminological point about the correct use of the term 'belief', since the case can be described in more neutral terms by employing David Chalmers' (ms) *subscript strategy*. Let us say that Peacocke's academic *believes$_1$* that foreign degrees are equal, although she *believes$_2$* that foreign degrees are inferior. We can add that her belief$_1$ is phenomenally individuated by its disposition to cause phenomenally conscious states of judgement, whereas her belief$_2$ is individuated by its disposition to cause actions, which conflict with her best judgements. Now we can ask whether or not Peacocke's academic has introspective access to these mental states. Intuitively, she has introspective justification to believe that she believes$_1$ that foreign degrees are equal, since this is what she is disposed to judge.[31] However, she does not have introspective justification to believe that she believes$_2$ that foreign degrees are inferior, since she has no disposition to judge this. If she knows what she believes$_2$, then she knows in the same way that others do – namely, by inference to the best explanation.[32] Therefore, whether we are restrictive or permissive in our use of the term 'believes', we may conclude that a mental state is introspectively accessible if and only if it is phenomenally individuated.

My proposal relies on the assumption that the phenomenology of judgement is proprietary, distinctive, and individuative. But why should we accept this controversial

assumption? David Pitt (2004) argues that we need to make this assumption in order to explain our introspective access to our own phenomenally conscious thoughts and judgements.[33] We can restate his argument as follows:

(1) It is possible to know by introspection what one is thinking and judging.

(2) If it is possible to know by introspection what one is thinking and judging, then one's thoughts and judgements have proprietary, distinctive, and individuative phenomenology.

(3) Therefore, one's thoughts and judgements have proprietary, distinctive, and individuative phenomenology.

The first premise is uncontroversial, while the second premise is motivated by generalization from other cases. For instance, it is plausible that the phenomenology of perceptual experience explains how introspective knowledge of perceptual experience is possible. By analogy, it is plausible that the phenomenology of judgement explains how introspective knowledge of judgement is possible.

The problem with Pitt's argument is that the analogy breaks down in the case of belief, since we have introspective knowledge of belief, but there is no phenomenology of belief that explains how introspective knowledge is possible. Therefore, the second premise cannot be motivated by a general principle to the effect that introspective knowledge of attitudes is possible only if those attitudes have proprietary, distinctive, and individuative phenomenology. Moreover, this raises an objection to the second premise, since it remains to be seen whether introspective knowledge of judgement should be explained on the model of perceptual experience, rather than the model of belief. However, if the arguments of this section are successful, then we should understand introspective knowledge of belief in terms of introspective knowledge of judgement, rather than vice versa. What I have argued, in effect, is that introspective knowledge of belief is

21

possible if and only if belief is phenomenally individuated by its disposition to cause the proprietary, distinctive, and individuative phenomenology of judgement. This raises a challenge for the opponents of cognitive phenomenology. If judgements do not have phenomenology that is proprietary, distinctive, and individuative, then what explains why one's beliefs are introspectively accessible, whereas one's subdoxastic states are not?

4. Some Problem Cases

I have argued that a mental state is introspectively accessible if and only if it is phenomenally individuated. We can formulate this thesis more precisely as follows:

· **The phenomenal accessibility thesis:** for all mental states M, necessarily, M is phenomenally individuated if and only if the following condition holds: one is in M if and only if one has introspective justification to believe that one is in M and one thereby has an introspective way of knowing that one is in M.

In this section, I will provide further support for the phenomenal accessibility thesis by defending it against a range of objections.

First, one might deny that the subject matter of introspection is restricted to one's *mental states*, as opposed to one's *physical states*. On this view, one has introspective justification to believe various propositions about one's own body. For instance, the experience of proprioception might provide introspective justification to believe that one's legs are crossed and the experience of agency might provide introspective justification to believe that one is raising one's arm. If so, then in cases of illusion, one has introspective justification to believe that one's legs are crossed, although it merely seems as if one's legs are crossed; or one has introspective justification to believe that one is raising one's arm, although it merely seems as if one is raising

one's arm. Therefore, we face counterexamples to the infallibility thesis in which one has introspective justification to believe false propositions.

On the simple theory, the experience of proprioception or agency does not provide introspective justification to believe propositions about one's own body. Certainly, one's justification is *peculiar* in the sense that it is not available to others, but it is not *privileged* in the way that is characteristic of introspective justification, since the source of one's justification is not identical with its subject matter. After all, it is not the case that one has justification to believe that one's body is in a certain state just by virtue of its being in that state. Rather, one has justification to believe that one's body is in a certain state by virtue of having an experience, which represents that the world is in that state. In cases of illusion, one's experience misrepresents the world and so one has justification to believe falsely that the world is in that state. As such, these are cases of perceptual justification, rather than introspective justification.

On the simple theory, one has introspective justification to believe propositions about one's experiences of proprioception and agency. For instance, one has justification to believe that it seems as if one's legs are crossed or that it seems as if one is raising one's arms just by virtue of the fact that it seems that way. Arguably, one also has non-introspective justification – perhaps a priori justification – to believe that one's experience represents veridically and hence that things are the way they seem. On this view, one has fallible justification to believe that one is in bodily state B, which derives from one's infallible introspective justification to believe that one is in some phenomenally individuated mental state M together with one's fallible non-introspective justification to believe that if one is in M, then one is in B. Therefore, one's justification depends partly, but not wholly, upon introspection.[34]

A similar response can be given to those who deny that the subject matter of introspection is restricted to *phenomenally individuated* mental states, as opposed to *externally individuated* mental states. The phenomenology of visual experience does not provide fallible introspective justification to believe that one sees a cup or that one sees that the cup is white. Rather, one has introspective justification to believe that one is in a certain kind of phenomenally individuated mental state – namely, it seems as if one sees a cup or it seems as if one sees that the cup is white. Moreover, one has non-introspective justification to believe that one's experience represents veridically and hence that things are as they seem. In other words, one's fallible justification to believe that one is in some externally individuated mental state M* derives from one's infallible introspective justification to believe that one is in some phenomenally individuated mental state M together with one's fallible non-introspective justification to believe that if one is in M, then one is in M*. Again, one's justification depends partly, but not wholly, upon introspection.

On this view, the phenomenology of judgement cannot provide introspective justification to believe that one is judging the externally individuated content that water is wet. Rather, one has introspective justification to believe that one is judging some phenomenally individuated content – for instance, that watery stuff is wet – but one also has non-introspective justification to believe that if one is thinking that watery stuff is wet, then one is thinking that water is wet. According to a phenomenal conception of justification, which propositions one has justification to believe is determined by one's phenomenally individuated mental states. If so, then my phenomenal duplicate on twin earth has justification to believe that he is thinking that water is wet, which is false, since he is not thinking about water, but twin water. As before, however, this fallible justification derives from his infallible introspective justification to believe that he is

thinking that watery stuff is wet and his fallible non-introspective justification to believe that if

he is thinking that watery stuff is wet, then he is thinking that water is wet. So, this case does not

provide a counterexample to the infallibility of introspective justification.

Nicholas Silins (this volume) argues that there are counterexamples to the infallibility of

introspective justification in which one judges a proposition and so one has introspective

justification to believe that one believes it, although one does not believe it because one's

judgement fails to manifest a stable disposition of the right kind. Here too, I claim that one's

justification depends partly, but not wholly, upon introspection. If one judges a proposition, then

one has introspective justification to believe that one judges it. Moreover, one has non-

introspective justification – perhaps a priori justification – to believe that if one judges a

proposition, then one believes it. So, if one judges a proposition, then one has fallible

justification to believe that one believes it, which derives from one's infallible introspective

justification to believe that one judges it together with one's fallible non-introspective

justification to believe that if one judges it, then one believes it. However, one has introspective

justification to believe that one believes a proposition if and only if one believes it.[35]

5. The Explanatory Question

In sections three and four, I defended an answer to the *generalization question*: a mental state is

introspectively accessible if and only if it is phenomenally individuated. This raises a version of

the *explanatory question*: why are phenomenally individuated mental states introspectively

accessible, rather than others, or none at all?

In order to explain which mental states are introspectively accessible, we need to invoke

distinctive features of those mental states together with more general claims in the theory of

justification, which explain why all and only mental states with those distinctive features are introspectively accessible. In this section, I consider and reject two *reductionist* strategies – namely, reliabilism and phenomenal conservatism – which attempt to answer the explanatory question by assimilating introspective justification to a more general model that applies equally to other non-introspective kinds of justification. The simple theory, by contrast, is *primitivist* in the sense that it regards introspective justification as a primitive or sui generis form of justification, which cannot be assimilated to a more general model that applies elsewhere. In the next section, however, I will develop an alternative strategy for answering the explanatory question that is consistent with the primitivism of the simple theory.

According to reliabilism, which propositions one has justification to believe is explained by the reliability of one's doxastic mechanisms.[36] What explains why one has justification to believe a proposition is that one has a doxastic mechanism, which is reliably disposed to yield belief in that proposition if and only if that proposition is true. In particular, what explains why one has introspective justification to believe a proposition is the fact that one has an introspective mechanism, which is reliably disposed to yield belief that one is in a mental state if and only if one is in that mental state.[37] On this view, a mental state provides introspective justification to believe that one is in that mental state if and only if it is reliably tracked by an introspective mechanism. Therefore, an answer to the explanatory question takes the following form:

(1) A mental state provides introspective justification to believe that one is in that mental state if and only if it is reliably tracked by an introspective mechanism.

(2) A mental state is reliably tracked by an introspective mechanism if and only if it is has feature F.

(3) Therefore, a mental state provides introspective justification to believe that one is in that mental state if and only if it has feature F.

However, this reliabilist answer to the explanatory question faces several problems.

First, the existence of a reliable introspective mechanism is not sufficient to explain one's introspective justification to form beliefs about one's mental states, since there are well known counterexamples to the claim that reliability is sufficient for justification. For instance, Norman has a reliable clairvoyant faculty, which reliably causes him to form true beliefs about the location of the President; Mr. Truetemp has a tempucomp implanted in his brain, which reliably causes him to form true beliefs about the temperature; the super-blindsighter has unconscious visual information, which reliably causes him to form true beliefs about objects in his blind hemi-field.[38] In each of these cases, the subject's beliefs are unjustified, but they are formed on the basis of a doxastic mechanism that is in fact reliable, although the subject in question does not know or have justification to believe that it is reliable. However, if mere reliability is not sufficient for justification in these cases, then why suppose that mere reliability about one's own mental states is sufficient for introspective justification?

Second, reliabilism is not sufficient to explain why introspective justification is infallible and self-intimated. If one's introspective mechanisms are reliable, but not infallible, then there will be cases in which one has introspective justification to believe that one is in a certain mental state, although one is not in that mental state. Similarly, if one's introspective mechanisms are reliable, but one's mental states are not self-intimating, then there will be cases in which one is in a certain mental state, although one does not have introspective justification to believe that one is in that mental state. In order to rule out these possibilities, introspective justification must be taken to require not just the usual degree of reliability, but the highest degree of reliability, which

is sufficient to ensure infallibility and self-intimation. However, reliabilism provides no principled motivation for imposing such a demanding requirement on introspective justification.

Third, reliabilism is not sufficient to explain the modal status of the introspective accessibility thesis. Reliabilism offers no modal guarantee that one has introspective justification to believe that one is in a certain mental state if and only if one is in that mental state. On the contrary, whether or not one has introspective justification to believe that one is in a certain mental state depends entirely on contingent empirical facts about one's psychology – namely, whether or not one has a reliable introspective mechanism of the relevant kind. According to reliabilism, then, it is a purely contingent matter which mental states, if any, provide one with introspective justification to believe that one is in those mental states. Therefore, we must look elsewhere in order to find a theory of justification that is capable of explaining the introspective accessibility thesis.[39]

According to *phenomenal conservatism*, which propositions one has justification to believe is explained by one's non-doxastic, phenomenal seemings.[40] What explains why one has justification to believe a proposition is the fact that it seems to be true and it is undefeated by any contrary seemings. On this view, seemings are distinct from beliefs: in some cases, a proposition seems true although one does not believe it. Moreover, seemings are distinct from conscious inclinations to believe: in many cases, it is because a proposition seems true that one has a conscious inclination to believe it. Thus, seemings may be said to comprise a sui generis category of conscious states, although they come in many different varieties, including perceptual, memorial and intellectual seemings.

Phenomenal conservatism motivates a version of the inner sense theory on which introspective justification is explained by reference to a special category of *introspective*

seemings.[41] On this view, one has introspective justification to believe that one is in a certain mental state if and only if it introspectively seems that one is in that mental state. Therefore, an answer to the explanatory question takes the following form:

(1) A mental state provides introspective justification to believe that one is in that mental state if and only if it introspectively seems that one is in that mental state.

(2) A mental state is such that it introspectively seems that one is in that mental state if and only if it has feature F.

(3) Therefore, a mental state provides introspective justification to believe that one is in that mental state if and only if it has feature F.

However, this answer to the explanatory question also faces several problems.

First, introspective seemings are epistemologically redundant. Consider my belief that I am in pain. If my belief is to be justified, then there must be some conscious state, which is distinct from my belief that I am in pain, but which explains and justifies my belief that I am in pain. However, this conscious state is just my pain. Why should we suppose that there is any further state of introspective seeming, which is distinct from my pain, but which explains and justifies my belief that I am in pain? Why isn't my pain sufficient by itself to explain and justify my belief that I am in pain?

Second, introspective seemings are phenomenologically redundant. According to some theories of phenomenal consciousness, a mental state is phenomenally conscious if and only if it seems that one is in it.[42] Arguably, however, this is either trivial or false. There is a trivial sense in which I am in pain if and only if seems to me that I am in pain. After all, pain is an experience; moreover, I experience my experiences in the same way that I smile my smiles and dance my dances. However, this does not imply that I represent my own experiences in the way

29

that my experiences represent the world. There are no obviously compelling grounds to suppose that I represent my own experiences; indeed, there are some compelling phenomenological grounds to suppose otherwise.[43]

Third, if we assume that all and only phenomenally conscious states are accompanied by introspective seemings, then phenomenal conservatism generates an overly restrictive answer to the generalization question. After all, beliefs are not phenomenally conscious states and so they are not accompanied by introspective seemings. Nevertheless, as I argued in section three, the fact that one believes a proposition is sufficient to provide introspective justification to believe that one believes it. Therefore, we must look elsewhere for an answer to the explanatory question that generates a more plausible answer to the generalization question.

What these explanatory strategies have in common is the reductionist ambition to explain which mental states provide introspective justification by appealing to a more general theory of justification, which applies equally to introspective and non-introspective forms of justification. According to the simple theory, by contrast, introspective justification is primitive or sui generis in the sense that it cannot be assimilated to a more general theory of justification that applies elsewhere. Introspective justification is the distinctive kind of justification that one has to believe that one is in a certain mental state, which one has just by virtue of being in that mental state. But if introspective justification is primitive or sui generis in this sense, then what resources do we have to answer the explanatory question? How can we explain why some mental states provide introspective justification, rather than others, or none at all?

My aim in the following section is to develop an answer to the explanatory question that is consistent with the primitivism of the simple theory. My strategy is to argue for the conditional claim that if access internalism is true, then the accessibility of facts about justification is best

explained by the *introspective accessibility constraint*, which states that the determinants of justification are introspectively accessible facts about one's mental states. If the introspective accessibility constraint is true, then we can answer the explanatory question as follows:

(1) A mental state provides introspective justification to believe that one is in that mental state if and only if it is among the determinants of justification.

(2) A mental state is among the determinants of justification if and only if it has feature F.

(3) Therefore, a mental state provides introspective justification to believe that one is in that mental state if and only if it has feature F.

Thus, we can explain our answer to the generalization question by appealing to a more general account of the determinants of justification. Alternatively, we can derive a more general account of the determinants of justification by appealing to our answer to the generalization question.

6. Access Internalism and Mentalism

One dimension of the debate between internalism and externalism in epistemology concerns the nature and extent of one's epistemic access to facts about which propositions one has justification to believe. Roughly speaking, *access internalism* is the thesis that one has a special kind of epistemic access to facts about which propositions one has justification to believe, whereas access externalism is the denial of access internalism. For current purposes, access internalism can be defined more precisely as the thesis that justification is accessible in the following sense:

· **Access internalism:** necessarily, one has justification to believe that *p* if and only if one has justification to believe that one has justification to believe that *p*.

Access internalism, like the accessibility thesis discussed in section two, is a biconditional that conjoins a *self-intimation* thesis, which is the conditional that goes from left to right, and an *infallibility* thesis, which is the conditional that goes from right to left. The self-intimation thesis rules out the possibility of having justification to be ignorant about which propositions one has justification to believe, whereas the infallibility thesis rules out the possibility of having justification to be in error. In other words, ideal rationality, according to access internalism, involves omniscience and infallibility about which propositions one has justification to believe.

Access internalism is a controversial thesis. For instance, reliabilism offers no modal guarantee that if one's first-order doxastic mechanisms are reliable, then one's second-order doxastic mechanisms are also reliable. So, if reliabilism is true, then access internalism is false, since it is possible that one has first-order justification to believe a proposition, although one lacks second-order justification to believe that one does.

Why should anyone suppose that access internalism is true? After all, it is not true in general that a condition C is accessible in the sense that C obtains if and only if one has justification to believe that C obtains. So, why suppose that justificatory conditions are any different? Here, I will briefly sketch three lines of argument, which I develop in much more detail elsewhere. First, access internalism explains intuitions about cases. Second, access internalism solves an epistemic version of Moore's paradox. And third, access internalism explains why justification is worth caring about, since it plays an important role in epistemic evaluation as an ideal of critical reflection.[44]

First, access internalism explains intuitions about cases, including clairvoyance and envatment.[45] Intuitively, my envatted duplicate has justification to believe propositions on the basis of perceptual experience, just as I do, although his beliefs are unreliable, whereas mine are

reliable. But what explains this intuition? My envatted duplicate has second-order justification to believe that he has justification to believe propositions on the basis of perceptual experience. So, by the infallibility thesis, he has first-order justification to believe propositions in this way. Conversely, my clairvoyant duplicate lacks justification to believe propositions on the basis of blind hunches or wishful thinking, just as I do, although his beliefs are reliable, whereas mine are unreliable. But what explains this intuition? My clairvoyant twin lacks second-order justification to believe that he has justification to believe propositions on the basis of blind hunches or wishful thinking. So, by the self-intimation thesis in its contrapositive form, he lacks first-order justification to believe propositions in this way.

Second, access internalism solves an epistemic version of Moore's paradox, since it explains what is wrong with believing Moorean conjunctions of the following forms:

(1) *p* and it is not the case that I have justification to believe that *p*.

(2) *p* and it is an open question whether or not I have justification to believe that *p*.

(3) I have justification to believe that *p* and it is not the case that *p*.

(4) I have justification to believe that *p* and it is an open question whether or not *p*.

If access internalism is true, then one cannot have justification to believe these Moorean conjunctions because one cannot have justification to believe each conjunct simultaneously. If access internalism is false, on the other hand, then it is left open that one can have justification to believe Moorean conjunctions. If self-intimation is false, then one can have justification to believe a proposition, although one lacks higher-order justification to believe that one has justification to believe it. In that case, one has justification to disbelieve or to withhold belief that one has justification to believe the proposition in question, so one has justification to believe either (1) or (2). Similarly, if infallibility is false, then one can have higher-order justification to

believe that one has justification to believe a proposition, although one lacks justification to believe it. In that case, one has justification either to disbelieve or to withhold belief in the proposition in question, so one has justification to believe either (3) or (4). Therefore, access internalism is needed in order to explain what is wrong with believing these Moorean conjunctions.

Finally, access internalism explains why justification is an important epistemic property that is worth caring about. Justification plays an important role in epistemic evaluation as an ideal of critical reflection: it is the epistemic property in virtue of which a belief has what it takes to survive ideal critical reflection. Roughly, one has justification to believe a proposition if and only if one would believe it after ideal critical reflection.[46] And yet one would not believe a proposition after ideal critical reflection unless one has second-order justification to believe that one has first-order justification to believe it. After all, the whole point of critical reflection is to bring one's beliefs into line with one's reflections about which propositions one has justification to believe. Therefore, ideal critical reflection involves believing a proposition if and only if one has second-order justification to believe that one has first-order justification to believe it. Hence, the role of justification in epistemic evaluation as an ideal of critical reflection provides the basis of an argument that justification is accessible:

(1) Necessarily, one has justification to believe that p if and only if one would believe that p after ideal critical reflection.

(2) Necessarily, one would believe that p after ideal critical reflection if and only if one has justification to believe that has justification to believe that p.

(3) Necessarily, one has justification to believe that p if and only if one has justification to believe that one has justification to believe that p.

In sum, there are compelling arguments for access internalism. Still, it is one thing to argue that access internalism is true, but it is another thing to explain what makes it the case that it is true. If access internalism is true, then part of the job description for a theory of justification is to give an account of the determinants of justification, which explains why access internalism is true.

There are no brute facts about which propositions one has justification to believe. On the contrary, these epistemic facts are *determined* by non-epistemic facts in the sense that no two situations can differ with respect to which propositions one has justification to believe unless those epistemic differences are explained by, or grounded in, corresponding non-epistemic differences. The *determinants* of justification are the non-epistemic facts (whatever they are) that determine the epistemic facts about which propositions one has justification to believe.

Arguably, the epistemic facts about which propositions one has justification to believe are determined *a priori* by non-epistemic facts. In other words, there are a priori conditionals of the form: necessarily, if such-and-such non-epistemic facts obtain, then such-and-such epistemic facts obtain. This is supported by the method of cases, in which a possible case is specified in non-epistemic terms and we can read off an epistemic specification without relying on any further empirical information. Given sufficient information about a case, it is a priori which propositions one has justification to believe in that case. If any further empirical information about the case is needed, then it can be included in a specification of the determinants of justification in which case the determination relation itself is a priori.[47]

One of the main tasks for a theory of justification is to specify the determinants of justification. According to reliabilism, for instance, the determinants of justification are non-epistemic facts about the reliability of one's doxastic methods. However, access internalism constrains an account of the determinants of justification, since justification is accessible only if

the determinants of justification are themselves accessible.[48] For instance, access internalism rules out a reliabilist account of the determinants of justification, since facts about the reliability of one's doxastic mechanisms are not accessible in the sense that one has justification to believe that they are reliable if and only if they are in fact reliable. But then what *are* the determinants of justification? Which facts, if any, are accessible in the sense that one has justification to believe that they obtain if and only if they obtain?

Access internalism provides the basis of an argument for a version of *mentalism*, according to which the determinants of justification are introspectively accessible facts about one's mental states.[49] The argument is as follows:

(1) Justification is accessible.

(2) Justification is accessible only if the determinants of justification are introspectively accessible facts about one's mental states.

(3) So, the determinants of justification are introspectively accessible facts about one's mental states.

Premise (2) is motivated by inference to the best explanation. If the determinants of justification are introspectively accessible facts about one's mental states and facts about the determination relation are accessible by a priori reflection, then facts about which propositions one has justification to believe are accessible by means of a combination of introspection and a priori reflection. Moreover, there is no other plausible explanation available. Therefore, access internalism stands or falls with the thesis that the determinants of justification are introspectively accessible facts about one's mental states.

Here, in more detail, is how the accessibility of justification is to be explained. Suppose that one has justification to believe that *p* in virtue of the fact that one is in some mental state M.

If M is introspectively accessible, then one has introspective justification to believe that one is in M. And if the determination relation is a priori accessible, then one has a priori justification to believe that if one is in M, then one has justification to believe that *p*. So, if one has justification to believe that *p* in virtue of being in M, then one has justification through a combination of introspection and a priori reflection to believe that one has justification to believe that *p* in virtue of being in M.[50] If, on the other hand, M is not introspectively accessible, then it is not the case that if one is in M, then one has introspective justification to believe that one is in M. So, it is not the case that, if one has justification to believe that *p* in virtue of being in M, then one has justification on the basis of introspection and a priori reflection to believe that one does.

In conclusion, if access internalism is true, then so is the introspective accessibility constraint, which states that the determinants of justification are introspectively accessible facts about one's mental states. Of course, it is a further question *which* facts about one's mental states are introspectively accessible. If the introspective accessibility constraint is true, however, then the answer to this generalization question constrains and is constrained by a more general account of the determinants of justification.

This provides a strategy for answering the explanatory question. We can explain our answer to the generalization question by appealing to a more general theory of the determinants of justification. For instance, we can explain the phenomenal accessibility thesis by appealing to a phenomenal version of mentalism, according to which the determinants of justification are phenomenally individuated mental states.[51] Alternatively, we can motivate a more general theory of the determinants of justification by appealing to an independently motivated answer to the generalization question. For instance, we can argue for a phenomenal version of mentalism by appealing to the phenomenal accessibility thesis. Moreover, we need not choose between these

options, since coherentism is more plausible than foundationalism as an epistemology for philosophy. In conclusion, the following package of claims is best regarded as part of a coherent and mutually reinforcing theory of justification:

(1) **The introspective accessibility constraint:** A mental state is introspectively accessible if and only if it is among the determinants of justification.

(2) **Phenomenal mentalism:** A mental state is among the determinants of justification if and only if it is phenomenally individuated.

(3) **The phenomenal accessibility thesis:** A mental state is introspectively accessible if and only if it is phenomenally individuated.

7. Rationality and Self-Knowledge

What is the connection between rationality and introspective self-knowledge? According to the simple theory, there is a necessary connection between rationality and self-knowledge. If one is ideally rational, then one is omniscient and infallible about one's phenomenally individuated mental states. In other words, introspective self-knowledge is a constitutive ideal of rationality.

A recurring theme in the literature on self-knowledge is that there is a necessary connection between rationality and self-knowledge.[52] And yet the existence of such a connection stands in need of further explanation. Why does ideal rationality require one to be omniscient and infallible about one's own mental states? Why can't one be ideally rational and yet introspectively blind to one's own mental states, just as one can be ideally rational and yet visually blind to the external world?

My aim in this paper has been to argue that the rational ideal of introspective self-knowledge is best explained as a consequence of access internalism in the theory of justification.

If justification is accessible in the sense that one has justification to believe a proposition if and only if one has higher-order justification to believe that one does, then ideal rationality involves omniscience and infallibility about which propositions one has justification to believe. Moreover, I argued that justification is accessible in this sense only if the determinants of justification are phenomenally individuated mental states, which are introspectively accessible in the sense that one has introspective justification to believe that one is in a phenomenally individuated mental state if and only if one is in that mental state. Therefore, if justification is accessible, then ideal rationality involves omniscience and infallibility about one's phenomenally individuated mental states, which determine which propositions one has justification to believe. Thus, the connection between rationality and self-knowledge is explained as a consequence of access internalism in the theory of justification combined with the claim that the determinants of justification are introspectively accessible facts about one's mental states.[53]

References

Armstrong, David. 1968. *A Materialist Theory of the Mind*. London: Routledge.

Block, Ned. 1997. "On a Confusion About a Function of Consciousness." *The Nature of Consciousness*, edited by N. Block, O. Flanagan and G. Guzeldere. Cambridge, MA: MIT Press.

Bonjour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

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. 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.

Chalmers, David and Frank Jackson. 2001. "Conceptual Analysis and Reductive Explanation." *Philosophical Review* 110: 315-61.

Chalmers, David. 2003. "The Content and Epistemology of Phenomenal Belief." *Consciousness: New Philosophical Perspectives*, edited by Q. Smith and A. Jokic. Oxford: Oxford University Press.

Chalmers, David. 2004. "The Representational Character of Experience." *The Future For Philosophy*, edited by B. Leiter. Oxford: Oxford University Press.

Chalmers, David. Forthcoming. "Verbal Disputes and Philosophical Progress." *Philosophical Review*.

Chomsky, Noam. 1957. *Syntactic Structures*. Hague: Mouton.

Cohen, Stewart. 1984. "Justification and Truth." *Philosophical Studies* 46: 279-95.

Conee, Earl and Richard Feldman. 2001. "Internalism Defended." *American Philosophical Quarterly* 38: 1-18.

Davies, Martin. 1987. "Tacit Knowledge and Semantic Theory: Does a Five Percent Difference Matter?" *Mind* 96: 441-62.

Davies, Martin. 1989. "Tacit Knowledge and Subdoxastic States." *Reflections on Chomsky*, edited by A. George. Oxford: Blackwell.

Evans, Gareth. 1981. "Semantic Structure and Tacit Knowledge." *Wittgenstein: To Follow a Rule*, edited by S. Holtzman and C. Leich. London: Routledge and Kegan Paul.

Evans, Gareth. 1982. *The Varieties of Reference*, Oxford: Oxford University Press.

Gertler, Brie. 2001. "Introspecting Phenomenal States." *Philosophy and Phenomenological Research* 63 (2): 305-28.

Gendler, Tamar. 2008. "Alief and Belief" *Journal of Philosophy* 105: 634-663.

Goldman, Alvin. 1979. "What is Justified Belief?" *Justification and Knowledge*, edited by G. Pappas. Dordrecht: Reidel.

Goldman, Alvin. 1993. "The Psychology of Folk Psychology." *Behavioral and Brain Sciences* 16: 15-28.

Horgan, Terence and John Tienson. 2002. "The Intentionality of Phenomenology and the Phenomenology of Intentionality." *Philosophy of Mind: Classical and Contemporary Readings*, edited by D. Chalmers. Oxford: Oxford University Press.

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

Huemer, Michael. 2001. *Skepticism and the Veil of Perception*. Lanham, MD: Rowman and Littlefield.

Huemer, Michael. 2007. "Compassionate Phenomenal Conservatism." *Philosophy and Phenomenological Research* 74 (1): 30-55.

Kriegel, Uriah. 2009. "Self-Representationalism and Phenomenology." *Philosophical Studies* 143: 357-81.

Lehrer, Keith. 1990. *Theory of Knowledge*. London: Routledge.

Locke, Don. 1967. *Perception and Our Knowledge of the External World*. London: Allen and Unwin.

Lycan, William. 1997. "Consciousness as Internal Monitoring." *The Nature of Consciousness*, edited by N. Block, O. Flanagan and G. Guzeldere. Cambridge, MA: MIT Press.

Marr, David. 1982. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. New York: Freeman.

Neta, Ram. 2011. "The Nature and Reach of Privileged Access." *Self-Knowledge*, edited by A. Hatzimoysis. New York, NY: Oxford University Press.

Nichols, Shaun and Stephen Stich. 2003. *Mindreading: An Integrated Account of Pretense, Self-Awareness and Understanding of Other Minds*. Oxford: Oxford University Press.

Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Peacocke, Christopher. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Peacocke, Christopher. 1998. "Conscious Attitudes, Attention and Self-Knowledge." *Knowing Our Own Minds*, edited by C. Wright, B. Smith and C. Macdonald. Oxford: Oxford University Press.

Pitt, David. 2004. "The Phenomenology of Cognition Or What Is It Like To Think That P?" *Philosophy and Phenomenological Research* 69 (1): 1-36.

Quine, Willard van Orman. 1970. "Methodological Reflections on Current Linguistic Theory." *Synthese* 21: 386-398.

Rorty, Richard. 1979. *Philosophy and the Mirror of Nature*. Princeton, NJ: Princeton University Press.

Rosenthal, David. 1997. "A Theory of Consciousness." *The Nature of Consciousness*, edited by N. Block, O. Flanagan and G. Guzeldere. Cambridge, MA: MIT Press.

Schwitzgebel, Eric. 2008. "The Unreliability of Naïve Introspection." *Philosophical Review* 117: 245-73.

Schwitzgebel, Eric. 2010. "Acting Contrary to Our Professed Beliefs, Or the Gulf Between Occurrent Judgment and Dispositional Belief." *Pacific Philosophical Quarterly* 91: 531-53.

Searle, John. 1990. "Consciousness, Explanatory Inversion and Cognitive Science." *Behavioural and Brain Sciences* 13: 585-96.

Shoemaker, Sydney. 1996. *The First Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Smithies, Declan. 2011. "What is the Role of Consciousness in Demonstrative Thought?" *Journal of Philosophy* 108 (1): 5-34.

Smithies, Declan. Forthcoming. "Moore's Paradox and the Accessibility of Justification" *Philosophy and Phenomenological Research*.

Smithies, Declan. Manuscript a. "Mentalism and Epistemic Transparency"

Smithies, Declan. Manuscript b. "Why Care About Justification?"

Smithies, Declan. Manuscript c. *The Epistemic Role of Consciousness*.

Sosa, Ernest. 2003. "Privileged Access." *Consciousness: New Philosophical Perspectives*, edited by Q. Smith and A Jokic. Oxford: Oxford University Press.

Siewert, Charles. 1998. *The Significance of Consciousness*. Princeton, NJ: Princeton University Press.

Steup, Matthias. 2009. "Are Mental States Luminous?" *Williamson on Knowledge*, edited by P. Greenough and D. Pritchard. Oxford: Oxford University Press.

Stich, Stephen. 1978. "Beliefs and Subdoxastic States." *Philosophy of Science* 45: 499-518.

Strawson, Galen. 1994. *Mental Reality*. Cambridge, MA: MIT Press.

Wedgwood, Ralph. 2002. "Internalism Explained." *Philosophy and Phenomenological Research* 65: 349-369.

Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.

Zagzebski, Linda. 1994. "The Inescapability of Gettier Problems." *Philosophical Quarterly* 44: 65-73.

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

Zimmerman, Aaron. 2007. "The Nature of Belief." *Journal of Consciousness Studies* 14: 61-82.

[1] Williamson (2000, Ch. 4) argues that no non-trivial conditions are luminous in the sense if they obtain, then one is in a position to know that they obtain. However, Smithies (manuscript, a) draws a distinction between epistemic and doxastic interpretations of luminosity and claims that Williamson's argument fails to show that there are no epistemically luminous, as opposed to doxastically luminous, conditions. See also the distinction between epistemic and doxastic versions of constitutivism in section two.

[2] Neta (2010) makes a related proposal, which he calls the *truth-sufficiency* account of privileged access. According to his account, S has privileged access to the fact that *p* if and only if *p* is itself a justification for S to believe that *p*.

[3] A prominent theme in the literature on self-knowledge is that rationality imposes limits on the possibility of ignorance and error about one's own mental states. For instance, Shoemaker (1996) argues against the possibility of *self-blindness* in which one is rational and conceptually competent, but one has no introspective knowledge of one's own mental states. Similarly, Burge (1996) argues against the possibility of *brute ignorance* and *brute error*, which is defined as ignorance and error that reflects no rational failure or malfunction in the subject.

[4] Compare Steup's (2009) discussion of the indefeasibility view.

[5] This recipe is proposed by Linda Zagzebski (1994, 69). She notes that the recipe applies only given some degree of modal independence between the justification condition for knowledge and the truth condition.

[6] Williamson's (2000, Ch. 4) anti-luminosity argument is sometimes regarded as a source of Gettier cases for introspectively justified beliefs, but see Smithies (manuscript, a) for arguments against this view.

[7] Here I am indebted to Daniel Stoljar who suggested these labels.

[8] This example is from Shoemaker (1996, 227). See also Armstrong's (1968, 93-4) long distance driver and Block's (1997, 386) unattended drill for further examples in a similar vein. The literature on inattentional blindness contains a further stock of examples.

[9] A version of this example appears in Locke (1967, 86). Gilbert Ryle's problem of the speckled hen gives rise to additional examples of misclassification; for further discussion, see Evans (1982, 228-9), Sosa (2003), and Smithies (manuscript a).

[10] For more detailed discussion and defence of this claim, see Smithies (forthcoming) and (manuscript, a & b).

[11] There may be weak doxastic conditions to the effect that one has justification to believe a proposition only if one has some doxastic capacities to form, maintain, and revise beliefs. If so, then the epistemic version of the accessibility thesis can be restricted to subjects that satisfy those weak doxastic conditions.

[12] Chalmers (2003) argues that all *direct* phenomenal beliefs are true, where a direct phenomenal belief is partly constituted by the phenomenal state that it picks out. Similarly, Horgan and Kriegel (2007) argue that all *bracketed* phenomenal beliefs are true, where a bracketed phenomenal belief is one that brackets out all relational information about the phenomenal state that it picks out.

[13] Compare Chalmers' (2003, 249) justification thesis: "When a subject forms a direct phenomenal belief based on a phenomenal quality, then that belief is prima facie justified by virtue of the subject's acquaintance with that quality."

[14] Compare Shoemaker, who writes, "if one has an available first-order belief, and has a certain degree of rationality, intelligence, and conceptual capacity (here including having the concept of belief and the concept of oneself), then automatically one has the corresponding second-order belief." (1996, xx)

[15] This may be what Shoemaker intends when he writes, "We should be wary of inferences from what can happen occasionally to what can happen as a matter of course: it may be true in Lake Wobegon that all of the children are above average, but it can't be true everywhere." (1996, 227)

[16] Peacocke (1992, Ch. 6) claims that one possesses the concept of belief only if one finds it primitively compelling to judge that one believes a proposition because one does in fact believe it.

[17] Shoemaker (1996, this volume) claims that second-order beliefs are partially constituted by the mental states they are about. Gertler (2001, this volume) claims that introspecting phenomenal states involves embedding the introspected state within an introspective belief. Chalmers (2003) claims that direct phenomenal beliefs are partially constituted by the phenomenal states they pick out. Horgan and Kriegel (2007) argue that phenomenal states are partially constituted by proto-beliefs, which are converted into bracketed phenomenal beliefs by the allocation of introspective attention.

[18] The simple view is therefore consistent with Schwitzgebel's (2008, this volume) claims about the unreliability of introspection and the absence of any dedicated mechanism of introspection.

[19] Stich (1978, 499) defines subdoxastic states as "psychological states that play a role in the proximate causal history of beliefs, though they are not beliefs themselves." We should add that these states are phenomenally unconscious in order to rule out perceptual experiences as examples of subdoxastic states.

[20] See Conee and Feldman (2001) for a discussion of the so-called 'problem of forgotten evidence'.

[21] Compare Zimmerman (2006, 357-61) and Shoemaker (this volume, xxx). See also Peacocke's so-called *NICS* cases, in which standing first-order beliefs justify occurrent second-order judgements in the absence of any intermediate conscious state: "Most of us, when it becomes conversationally appropriate to say 'I know my name is NN', or 'I know my address is such-and-such', have no need to wait upon its surfacing in consciousness what our names and addresses are. We make these utterances intentionally and knowledgeably, but not because it has just occurred to us that our names and addresses are such-and-such." (1998: 91)

[22] Shoemaker (this volume) appeals to availability, which is closely related to access consciousness. Compare Zimmerman, who writes: "It is at least *difficult* to imagine a subject who believes that p in a fully in a fully A-conscious manner – where this belief is poised to guide her inferences and behaviour in all the customary ways – and who is also caused to believe that she believes that p by the fact that she believes that p, but who is nevertheless unjustified in believing that she has this belief. And this lends considerable *prima facie* support to the claim that A-consciousness is sufficient for direct accessibility." (2006, 357)

[23] Higher-order theories of consciousness are proposed in various forms by Armstrong (1968), Lycan (1997), Rosenthal (1997) and Carruthers (2000).

[24] See Smithies (2011) for further discussion of the role of consciousness in perceptual justification.

[25] Individuation, as I am using the term, is closely tied to the notion of essence: what individuates a thing is what makes it what it is, but the essence of a thing is what makes it what it is, so the essence of a thing is what individuates it. For example, a set is individuated by its members in the sense that the essence of a set is exhausted by its members. So, to say that a mental state is phenomenally individuated is to say that its essence is exhausted by its relations to phenomenally conscious states.

[26] The point is familiar from the literature on Quine's (1970) challenge to Chomsky's (1959) notion of tacit knowledge; see, for instance, Evans (1981) and Davies (1987).

[27] Compare Strawson (1994, Ch. 1), Siewert (1998, Ch. 8), Horgan and Tienson (2002) and Pitt (2004).

[28] The phenomenology of judgement, like the phenomenology of perceptual experience, is specific with respect to narrow content, which is determined wholly by phenomenology, but not wide content, which is determined in part by relations to the environment. In my view, judgements are individuated by their narrow contents, but not their wide contents. See Horgan and Tienson (2002) and Chalmers (2004) for further discussion of the relationship between phenomenology and intentionality.

[29] Searle (1990) argues that beliefs are accessible to consciousness in the sense that they are potentially conscious. However, beliefs are not potentially conscious: they are dispositions to cause conscious states, which are distinct from their potentially conscious manifestations. The *contents* of beliefs are accessible to consciousness as the contents of judgements, but this is a consequence of the fact beliefs are individuated by the disposition to cause judgements with the very same contents.

[30] There are at least four possible descriptions of Peacocke's academic: (i) she believes that foreign degrees are inferior and does not believe that they are equal (Peacocke 1998); (ii) she believes that foreign degrees are equal and does not believe that they are inferior (Zimmerman 2007, Gendler 2008); (iii) she believes both (Shoemaker, this volume); or (iv) she does not determinately believe either (Schwitzgebel 2010). Only the second option does justice to the agent's epistemic rationality. See Zimmerman (2007) and Gendler (2008) for related discussion.

[31] What if the disposition is blocked by drunkenness or temporary insanity? I maintain that the belief is introspectively accessible in the sense that the subject has introspective justification to believe that she has this belief, although she may be unable to take advantage of her justification unless the blockage is removed. In that case, the blockage constitutes a departure from full rationality.

[32] As Brie Gertler pointed out to me, the inference may be based on introspective knowledge of the results of an imaginative simulation exercise, rather than observation of one's behaviour. In that case, according to the simple theory, the premises of the inference are known by introspection alone, but not the conclusion.

[33] Goldman (1993) gives a related style of epistemological argument for cognitive phenomenology.

[34] In the terminology of mediate versus immediate justification: one does not have immediate introspective justification to believe propositions about one's own body, since one's justification is mediated by non-introspective justification to believe other propositions. I remain neutral on the further question whether one's perceptual justification to believe propositions about one's own body is mediate or immediate.

[35] Silins (this volume, p. x, fn. 19) anticipates this reply and argues that it over-intellectualizes our ways of gaining introspective self-knowledge. The issue deserves further discussion, but it is not clear to me that analogous views in the epistemology of perception are committed to an over-intellectualization.

[36] I have formulated reliabilism in terms of doxastic mechanisms, but Goldman (1979) appeals to processes, Nozick (1981) appeals to methods and Sosa (this volume) appeals to competences. These differences in formulation are not crucial for current purposes.

[37] Examples of introspective mechanisms include Armstrong's (1968) self-scanning mechanism and Nichols and Stich's (2003: 160-4) monitoring mechanism.

[38] These examples are from BonJour (1985, 41), Lehrer (1990, 163-4) and Block (1997, 385-6).

[39] For related criticisms of reliabilism, see Shoemaker's (1996) arguments against the broad perceptual model and Burge's (1996) arguments against the simple observational model, which hinge on the claim that they cannot explain the limits on the possibility of ignorance and error about one's own mental states.

[40] Huemer's (2001, 99) rule of phenomenal conservatism states that: "If it seems to S as if P, then S thereby has at least prima facie justification for believing that P." Huemer (2007) also endorses the right-to-left conditional, which plays a role in his self-defeat argument for phenomenal conservatism.

[41] Huemer (2007, 30) includes "apparent introspective awareness" in his inventory of seemings.

[42] This assumption is often used as a premise in arguing for higher-order theories of consciousness. See Armstrong (1968), Lycan (1997), Rosenthal (1997), Carruthers (2000) and Kriegel (2009).

[43] See Shoemaker (1996, Ch. 10) and Siewert (this volume) for arguments against inner sense.

[44] See Smithies (forthcoming) and (manuscript, b). For simplicity, I focus here on arguments that the presence of justification is accessible, but similar arguments show that the absence of justification is also accessible.

[45] See BonJour (1985, Ch. 3) for the clairvoyance case and Cohen (1984) for the new evil demon case, which is a variation on the envatment case.

[46] This needs further refinement to deal with conditional fallacy objections. See Smithies (manuscript, b) for details.

[47] See Chalmers and Jackson (2001, section 2) for related discussion of a priori entailment relations.

[48] Strictly speaking, justification is accessible only if there are *some* accessible determinants of justification, although these may be determined themselves by inaccessible determinants of justification.

[49] Note that many epistemologists accept mentalism, but reject accessibilism, including Williamson (2000), Conee and Feldman (2001) and Wedgwood (2002).

[50] Here I assume that one's justification transmits across the entailment from (i) one is in M; and (ii) if one is in M, then one has justification to believe that p; to (iii) one has justification to believe that p.

[51] In this paper, I do not provide independent arguments for the phenomenal version of mentalism, although I do so elsewhere. See Smithies (manuscript, c).

[52] Classic discussions of the relationship between rationality and self-knowledge include Burge (1996) and Shoemaker (1996), but see also the chapters by Moran, Shoemaker and Stoljar (this volume) for further discussion.