Concepts, Conceptions and Self-Knowledge[1]

Sarah Sawyer

University of Sussex

Forthcoming in *Erkenntnis*

(This is a pre-proof version)

1. Introduction

Content externalism is the view that mental representation depends constitutively on non-representational (typically causal) relations between a subject and her wider environment. It has gained widespread support since its introduction into the mainstream philosophical literature by Burge, following externalist claims about the nature of language provided by Kripke and Putnam.[2] I do not argue for content externalism in this paper. Nor do I engage directly with the vast body of literature that addresses the question of whether content externalism is compatible with privileged access to thought-content or with the authoritative nature of self-ascriptions.[3] Rather, I take content externalism as my starting point and argue that two implications of the view require us to rethink the nature of self-knowledge. The first implication is a distinction between concepts and conceptions. The second implication is a distinction between thoughts (understood as propositional attitudes) and states of mind. Although the second distinction is not typically recognised, it follows from the first. Taking these implications on board, I provide an externalist account of self-knowledge which is shaped by them.

[1] This paper was written in connection with Project FFI2012-38908-C02-02: *Self-Knowledge, Expression and Transparency*, funded by MINECO, a branch of the Spanish government. I would like to thank members of the *Self-Knowledge, Expression and Transparency* research group, members of the LOGOS research group, and Tony Booth for comments on earlier drafts. I am especially grateful for the detailed and helpful comments provided by two anonymous referees for this journal. The paper has improved immeasurably as a result of their input.

[2] See Burge (1979, 1986) for the claim that mental content fails to supervene locally on the intrinsic properties of an individual thinker. See Kripke (1972, 1980) and Putnam (1973, 1975) for the claim that linguistic meaning fails to supervene locally on the intrinsic properties of an individual speaker.

[3] For a brief insight into this vast literature see for example the edited collections by Pessin and Goldberg (1996), Ludlow and Martin (1998) and, more recently, Goldberg (2015).

One theoretical advantage of the account is that it offers a principled way to navigate a middle path between the two extremes of absolute epistemic security on the one hand and scepticism about first-personal self-knowledge on the other.[4] The middle path accords with intuition. In our ordinary conversation, we offer spontaneous pronouncements about our own mental states: "I'm so happy you could make it"; "I wish the train strikes would come to an end"; "I think I'll go for a walk after lunch"; and we frequently inquire after the mental states of others, expecting them to have some insight into their own minds that is not immediately accessible to us: "What did you think of the lecture?"; "Do you want to meet later?"; "Do you intend to eat that last biscuit?". Self-ascriptions are, in the main, taken as authoritative because we are presumed to have privileged access to our own thoughts.[5] Nonetheless, self-ascriptions are, again in the main, taken to be open to behavioural counter-evidence. Jo may profess to like a colleague while nonetheless consistently avoiding his company and habitually complaining about his treatment of others. In such a case, Jo's behaviour provides third-personal counter-evidence to her first-personal self-ascription. The externalist account of self-knowledge I propose makes sense of these phenomena.

It is not uncommon for a theory of self-knowledge to aim to tread the middle path. However, those who accept the epistemic security of self-ascriptions tend to treat behavioural counter-evidence as anomalous, to be categorized as the result of the Freudian subconscious rather than as integral to a proper account of self-knowledge; and those who take behavioural evidence seriously tend to downplay the role and significance of epistemic security.[6] In contrast, my proposal offers an integrated account of first-personal self-knowledge within an externalist framework that explains both epistemic security and behavioural counter-evidence. It affords an authority to a subject's self-ascriptions when those ascriptions are based on privileged access to the thoughts ascribed, but it also acknowledges the fact that a subject's self-ascriptions are open to behavioural counter-evidence, and it does so because this possibility follows in a straightforward way from the distinction between thoughts and states of mind.

[4] For a traditional account that offers absolute epistemic security see Descartes (1641). For scepticism about substantial self-knowledge see Cassam (2014). See Shoemaker (1988) for arguments against the coherence of the sceptical view.

[5] This contrasts with the account found in Ryle (1949) according to which authority is grounded not in privileged access but in an abundance of behavioural evidence.

[6] For an example of the latter position see Schwitzgebel (2012).

The structure of the paper is as follows. In section 2, I introduce the distinction between concepts and conceptions and the consequent distinction between thoughts and states of mind. In section 3, I propose an externalist account of self-knowledge and discuss the preliminary epistemological implications of the proposal, relating it to questions of privilege and authority. In section 4, I extend and elaborate the account, explaining in more detail the implications of the distinction between thoughts and states of mind. In section 5, I relate the account to behavioural evidence and counter-evidence. In order to clarify the view, I draw comparisons between self-knowledge and perceptual knowledge throughout the paper. I conclude in section 6.

Note that the paper has two separable aims. The first is to show that content externalism requires us to rethink the nature of self-knowledge in light of the distinction between concepts and conceptions and the consequent distinction between thoughts and states of mind. The second is to offer an account of self-knowledge that accommodates the relevant distinctions. It may be possible for other accounts of self-knowledge to accommodate the distinctions, but that remains to be seen.

2. Concepts and Conceptions; Thoughts and States of Mind

I take concepts to be mental representations that are components of the propositional contents of thoughts, and I take the collection of beliefs a subject associates with a concept to be the subject's conception of the relevant subject matter. For example, my concept of penguins is the mental representation PENGUIN, and the conception I associate with the concept is that of flightless birds who are mostly black and white, that can identify their mate by sound, that come in different varieties, that only live in the Southern Hemisphere, and so on. Note that the conception has here been specified as a set of beliefs about the subject matter referred to by the associated concept, but without explicit mention of the associated concept itself; the term 'penguin' does not appear in the specification of the associated conception. Note also that both concepts and conceptions admit of a type-token distinction. Whether it is types or tokens that is relevant on a given occasion will be stated explicitly only when it is not clear from the context.

Content externalism maintains that mental representation depends constitutively on relations between a subject and her wider environment. It embodies the rejection of the

(internalist) claim that a subject's concepts are determined by her associated conceptions.[7] Content externalism thus entails not just a theoretical, but a substantive distinction between concepts and conceptions. Indeed, content externalism implies that possession of the same concept by two individuals is neither necessary nor sufficient for possession, by those individuals, of the same associated conception. Putnam's Twin Earth thought experiment demonstrates the failure of necessity. Oscar and Twin Oscar are assumed to associate the same conception with their different concepts WATER and TWIN WATER. The failure of necessity is also demonstrated by Burge's example of the subject who actually possesses the concept SOFA but counterfactually possesses the concept SAFO, since the subject's associated conception is assumed to remain constant across the actual and the counterfactual scenarios. Putnam's example of elms and beeches demonstrates the failure of sufficiency. Putnam and a tree expert are assumed to associate different conceptions with the concept ELM. The failure of sufficiency is also demonstrated by Burge's example of Alf, whose conception of arthritis differs from his doctor's.[8]

The substantive distinction between concepts and conceptions both implies that and depends on the fact that an individual can possess a concept even though her associated conception is vague or inaccurate. Again, the traditional arguments of Putnam and Burge demonstrate this fact. Putnam claims to possess the concept ELM, even though his associated conception is not precise enough to distinguish elms from beeches; and Alf is claimed to possess the concept ARTHRITIS even though his associated conception is, mistakenly, of a disease that can spread to the muscles. Thus grasp of a concept comes in degrees. The tree expert grasps the concept ELM more fully than Putnam; and Alf's doctor grasps the concept ARTHRITIS more fully than Alf. Mere possession of a concept is one thing; complete grasp of a concept is another. Crudely speaking, we can say that the better a subject's grasp of a concept, the more accurate her associated conception of the relevant subject matter will be. The claim is crude, because grasp of a concept is related not only to a subject's associated conception but also to her capacity for correct deployment of the concept across a range of appropriate contexts. Grasp of a concept will therefore vary across at least two dimensions. Nonetheless, since full grasp of a concept is an idealised absolute, we can say that full grasp

[7] Content internalism is defined by a local supervenience thesis according to which what a subject thinks is determined by her intrinsic (typically physical) states. The claim that concepts are determined by conceptions is just one form of content internalism.

[8] For the examples, see Putnam (1973, 1975) and Burge (1979, 1986).

of a concept will involve a correct associated conception together with a capacity for correct deployment under normal conditions across a range of contexts. It is against this idealised measure that we should understand the claim that the better a subject's grasp of a concept, the more accurate her associated conception of the relevant subject matter will be. On the flip side, an incomplete grasp of a concept reflects a degree of ignorance about the relevant subject matter, and this will manifest itself not only in false beliefs but in incorrect contextual applications.[9] It is for this reason that content externalism is best understood as a general thesis about the nature of mental representation *per se*, rather than as a thesis limited to natural kind concepts, or even to empirical concepts more broadly construed. The possibility of ignorance is not limited to specific topics of inquiry, and hence incomplete grasp of concepts is potentially universal.

The first implication of content externalism, then, is that there is a substantive distinction between concepts and conceptions. The second implication, which follows from the first, is that there is a distinction between thoughts, understood as propositional attitudes, and states of mind. If two people can possess the very same concept and yet have different associated conceptions of the relevant subject matter, then two people can have the same thought without being in the same state of mind. Thus Putnam and the tree expert may both believe that there are fewer elms in England than there used to be, but this is consistent with their being in different states of mind; and Alf and his doctor may both believe that arthritis is painful, but this is consistent with their being in different states of mind. This requires further explanation.

A thought is individuated by an attitudinal relation to a propositional content, and the propositional content of a thought is individuated by its constituent concepts. Putnam's thought is type-identical to the tree expert's thought because Putnam and the tree expert bear the same attitude to the same propositional content. Similarly, Alf's thought is type-identical

[9] It might be objected that two individuals can have different conceptions associated with a single concept even though neither is ignorant of the subject matter. For example, consider a knowledgeable pathologist and a knowledgeable microbiologist. Let us suppose that they each possess the concept BACTERIA, but have different associated conceptions because of their different fields of study. Wouldn't this be a case in which they have different conceptions but nonetheless each has a full grasp of the relevant concept? The answer is 'no'. The alleged counterexample rests on a confusion. While each of them is knowledgeable relative to their own field of study, each of them is also ignorant relative to the other's field of study. This means that they will each have an incomplete grasp of the concept BACTERIA despite their knowledge. Note that if they were knowledgeable about each other's field of study, it is hard to see how their conceptions would differ.

to his doctor's thought because Alf and his doctor bear the same attitude to the same propositional content. This makes the individuation of thoughts blind to the different ways in which the constituent concepts that individuate a thought's propositional content are grasped by the relevant individual. Nonetheless, an individual's grasp of the constituent concepts of her thought is clearly relevant to her psychological state. Putnam and the tree expert are not, I suggest, in the same psychological state when they each believe that there are fewer elms in England than there used to be; and Alf and his doctor are not in the same psychological state when they each believe that arthritis is painful. In order to capture this difference we need to appeal to a psychological phenomenon other than thoughts. This is the role I assign to states of mind. A subject's state of mind, as I will understand it, is determined not by her thought alone, but by her thought in conjunction with the way in which, or the extent to which, she grasps each of the constituent concepts in her thought's propositional content. Note that this means that there is a psychological difference between Putnam and the tree expert not just in the sense that their total sets of thoughts differ, but in the sense that there is a difference between them in the very act of thinking the thought that there are fewer elms in England than there used to be. Similarly, there is a psychological difference between Alf and his doctor not just in the sense that their total sets of thoughts differ, but in the sense that there is a difference between them in the very act of thinking the thought that arthritis is painful.

Of course, given that a conception is here understood as a set of beliefs, there will be a difference in the total set of beliefs, and hence the total set of thoughts, of two individuals whenever they associate different conceptions with a constituent concept of a particular thought. Nonetheless, it is possible for two individuals to grasp the constituent concepts of a particular thought in the very same way, and hence be in the same state of mind with respect to that thought, even though their total sets of thoughts differ. Thus a difference in a (particular) state of mind cannot be equated with a difference in a total set of thoughts or, as we might put it, with a difference in an 'overall' state of mind.

The distinction between thoughts and states of mind is required so long as we acknowledge that grasping a concept comes in degrees; it is implied, that is, by the distinction between concepts and conceptions. The fact that two individuals can have the same thought and yet be in different states of mind relative to that thought can be, although need not be,

articulated in terms of a contrastive account of the propositional attitudes.[10] According to a contrastive account of the propositional attitudes, a subject believes that *p* not *simpliciter*, but in a context, against certain background assumptions, and always in a way that is sensitive to the conceptions she associates with the concepts involved in her belief. Thus *S* may believe that *p* rather than that *q*, *r* or *s*, while *S'* believes that *p* rather than that *s*, *t* or *v*. To take an example, suppose Alf and his doctor both believe that arthritis occurs in the joints. For Alf's doctor, the proposition that arthritis occurs in the joints contrasts with the proposition that arthritis occurs in the muscles. Alf's doctor therefore *believes* that arthritis occurs in the joints *rather than* that arthritis occurs in the muscles. For Alf, the propositions are not contrasting propositions. As a result, Alf does *not* believe that arthritis occurs in the joints *rather than* that arthritis occurs in the muscles. Alf believes that arthritis occurs in the joints relative to a different set of contrasting propositions, including, for example, the proposition that arthritis occurs in the brain, the proposition that arthritis occurs in the liver, and so on, many of which propositions will also be in the set of contrasting propositions relative to which his doctor believes that arthritis occurs in the joints.

The relevant contrasting propositions that individuate a given subject's contrastive propositional attitude are generated by, and hence reflect, her grasp of the concepts that constitute its non-contrastive propositional content. The propositional content of a propositional attitude is non-contrastive in the sense that what is believed is, simply, that arthritis can occur in the joints. It is the relation to the propositional content that is contrastive. Alf and his doctor believe the same propositional content, but relative to different propositional contrast classes. And it is, ultimately, the fact that Alf and his doctor grasp the concept ARTHRITIS differently that explains their different (contrastive) propositional attitudes. A contrastive understanding of the propositional attitudes thus captures the fact that when two individuals believe that *p*, there is a sense in which they believe the same thing, but also, normally, a sense in which they do not believe the same thing. They have the same attitude towards the same propositional content, but their grasp of the concepts that constitute its propositional content will normally differ. The two distinct senses here are captured by the distinction between thoughts on the one hand, understood as propositional attitudes which are

[10] For a contrastive account of the propositional attitudes see Sawyer (2014). For a theoretical benefit of the view see Sawyer (2015). The distinction between thoughts and states of mind can also be articulated within a framework that acknowledges degrees of belief, although I will not provide the details here.

not, as yet, relativized to a contrast class of propositions, and states of mind on the other, understood as propositional attitudes relativized in the relevant way.

The having of a particular thought, then, is consistent with a subject's being in different states of mind. This is crucial to the account of self-knowledge that I propose in the next section because the distinction between thoughts and states of mind opens up the possibility that the correct ascription of a thought to a subject does not fully capture her state of mind. The correct ascription of a thought states the attitudinal relation between a thinker and a propositional content.[11] This is an important function of thought ascriptions because the correct ascription of a thought to a subject tells us what that subject is thinking about and how; concepts, being externally individuated, connect the thinker representationally to a subject matter.[12] As a result, however, the correct ascription of a thought does not reflect, or carry information about, the conceptions that the individual associates with the constituent concepts of the thought's propositional content. This means that the correct ascription of a thought to a subject reflects her deployment of the constituent concepts in thought, and hence her possession of those concepts, but it does not reflect the extent to which those concepts have been grasped by her; as such, it does not fully reflect her state of mind, but, rather, leaves it underdetermined.

3. The Partial-Representation Model of Self-Knowledge

In this section, I introduce an account of self-knowledge that is shaped by the two implications of content externalism explained above. I'll call the account the 'partial-representation model of self-knowledge'. In brief, the account maintains that the self-ascription of a thought is authoritative when it is based on, and hence derives its epistemic warrant from, a conscious thought in virtue of which it partially represents an underlying state of mind. The term 'conscious' is subject to multiple ambiguity. For present purposes, a thought is conscious, for a given subject, just in case it is a thought which the subject is currently thinking. Conscious thoughts in this sense are occurrent thoughts. In what follows I use the terms 'conscious thought', 'occurrent thought' and 'conscious, occurrent thought'

[11] For present purposes, I abstract away from the multiple complexities surrounding the practice of ascribing propositional attitudes and focus exclusively on the straightforward case in which the ascription specifies the thought ascribed.

[12] For more on the nature and significance of the connection between a concept and a subject matter see Sawyer (2018, 2019).

interchangeably. Authoritative self-ascriptions understood along these lines will be subject to behavioural counter-evidence because of the partial nature of the representation involved both at the level of the conscious thought and, hence, at the level of the self-ascription. In this section, I explain in basic terms how a conscious thought can warrant a self-ascription. In section 4, I explain the phenomenon of partial representation. In section 5, I discuss the role of behavioural counter-evidence in the account.

I start, then, with an example borrowed from Peacocke. The example is designed to show how a conscious thought can warrant a self-ascription, and hence provides a basic template for my own account of self-knowledge. Towards the end of this section, I examine two further examples from Peacocke, neither of which he thinks can be understood as cases of self-ascription warranted by conscious thought. My own account, by contrast, accommodates the examples within a single framework. Authoritative, first-personal self-knowledge can in general, I maintain, be understood as warranted by a conscious thought in virtue of which it partially represents an underlying state of mind. Here is Peacocke's first example.

(Case 1) The case of Napoleon

> 'Suppose you are asked 'Where was Napoleon defeated?' You try to remember; and memory serves up the information that Napoleon was defeated at Waterloo. It's occurring to you that he was defeated at Waterloo can be a subjective, conscious event which is capable… of engaging your attention. … Suppose you make the first-person self-ascription 'I believe that Napoleon was defeated at Waterloo', and make it because (a) you seem to remember that Napoleon was defeated there, and because (b) you are taking your memory at face value. … In cases of this type, the thinker is entitled to the self-ascription of the belief'. (Peacocke, 1996: 119-23)

Peacocke talks of the conscious state providing the subject with an *entitlement* for the self-ascription. The term 'entitlement' is used inconsistently in the literature[13], but here Peacocke's use is intended to indicate that the self-ascription is epistemically warranted

[13] Different uses of the key epistemic terms 'warrant', 'justification' and 'entitlement' is widespread, but for a sense of the differences with respect to the latter, see for example Burge (1993), Dretske (2000), Peacocke (1999) and Plantinga (1993). As will become clear, in the present context I adopt Burge's use.

independently of whether the subject can articulate the warrant, or provide an explanation that counts as a reason for the self-ascription. It is thus intended to avoid an over-intellectualized understanding of epistemic warrant according to which beliefs are warranted only if the subject can provide reasons (to another) in their favour.

Peacocke's use of the term has its roots in a distinction due to Burge between two kinds of epistemic warrant: justifications and entitlements. However, Peacocke's use of the term 'entitlement' does not exactly align with the distinction Burge draws. According to Burge, what differentiates justification and entitlement is, crudely, that justification involves propositional (that is, conceptual) states of the subject, whereas entitlement does not.[14] Burge says: 'Being *justified* requires having in one's psychology a reason that is operative', where 'a reason that one has for an attitude is *operative* if and only if the reason figures in a cognitively relevant causal way in forming or sustaining the attitude. … Being *entitled* to a belief is being warranted in holding it, without depending for being warranted on having an operative reason for it. Entitlement is warrant without reason.' (Burge 2013: 490, original emphasis).

The distinction between having a reason and having an operative reason is, effectively, the distinction between propositional justification and doxastic justification. Very roughly, a subject S has propositional justification for believing that *p* when some propositional states of hers provide evidential support for the belief that *p*, even if her belief that *p* is not caused by or based on those propositional states, and even if S does not believe that *p*. A subject has doxastic justification for her belief that *p*, in contrast, when she believes that *p* and her belief that *p* is caused by or based on the relevant propositional states. The notion of entitlement is different entirely, since it is warrant that is not propositional.

The distinction between justification and entitlement is drawn, then, to avoid an over-intellectualized, internalist understanding of epistemic warrants. But the point to which Burge wishes to draw attention is that while some beliefs are warranted because they are supported by propositional states the subject has, in which case they are justified, some beliefs are not supported by propositional states the subject has and yet are warranted nonetheless, and these are beliefs to which the subject is entitled. Perceptual beliefs provide the classic example of beliefs to which we are entitled. This is because a perceptual state can, under certain conditions, provide an epistemic warrant for a perceptual belief, even though the perceptual

[14] The distinction is unaffected by the distinction between concepts and conceptions with which I began. This is because concepts are elements of thoughts.

state is not propositional in form.[15] The distinction between justification and entitlement is, Burge says, a 'conceptual/grammatical' one. (Burge 2013: 490).

Taking Burge's distinction on board, I propose that a conscious, occurrent thought, which is propositional in form, is best understood as providing a *reason*, and hence a *justification* for a self-ascription, rather than as providing an *entitlement*. This way we can acknowledge the crucial epistemic distinction to which Burge draws attention—that between warrant provided by propositional states and warrant provided by non-propositional states—while also accommodating the point Peacocke intended to highlight—that a self-ascription can be warranted even though the subject is unable to explain or articulate her warrant. When a self-ascription is based on a conscious, occurrent thought, the conscious thought is an operative reason for her self-ascription. But even when a subject has an operative reason for a self-ascription—even when her self-ascription is epistemically grounded in her conscious thought—she may nonetheless not be in a position to provide a reason in the sense of articulating it to another; her warrant does not depend on her capacity to do so.

To say that a subject's warrant for an authoritative self-ascription does not depend on her capacity to articulate that warrant is not, of course, to say that she is precluded from being able to do so. A look at the perceptual case here is instructive. Suppose my perceptual experience as of a book on my desk entitles me to believe that there is a book on my desk. If someone were to ask me why I think there is a book on my desk, I might say 'because I can see it'. This explanation might contrast with other potential explanations, such as that my friend told me it was there, or that I remember putting it there this morning. The response I offer explains the belief as a perceptual belief but it does not warrant it; the perceptual belief is warranted even if I can say nothing. Now, by analogy, suppose my conscious thought that my aunt is kind justifies my self-ascription that I believe that my aunt is kind. If someone were to ask me why I think I believe my aunt is kind, I might say 'because that's what I think'. This explanation might contrast with other potential explanations, such as that my psychotherapist told me that's what I believe, or that I have figured it out from my behaviour. The response I offer explains the belief as an authoritative self-ascription but it does not warrant it: the self-ascription is warranted even if I can say nothing.

[15] If perceptual states were propositional, they would count as providing reasons for perceptual beliefs. This is, effectively, the view given in McDowell (1994). I leave this to one side here because the question at issue in the paper is whether conscious states should be classified as reasons, not whether perceptual states should be so classified.

On the partial-representation account of self-knowledge, then, conscious thoughts provide reasons for self-ascriptions which, when operative, provide epistemic grounds for the relevant self-ascriptions. This accounts for the epistemic status of self-ascriptions as authoritative. Self-ascriptions are authoritative because they are epistemically grounded in conscious, occurrent thoughts to which a subject has privileged access; and occurrent thoughts are ones to which a subject has privileged access because they are, of necessity, occurrent only for the thinker at the time. Conscious thoughts therefore provide objective but essentially first-personal reasons for self-ascriptions. This is analogous to the perceptual case in the sense that perceptual states provide objective but essentially first-personal entitlements for perceptual beliefs. A full account of how perceptual states provide entitlements for perceptual beliefs must make reference to the constitutive connection between the representational content of perceptual states, the representational content of perceptual beliefs and the content-determining environment. Content externalism therefore provides the requisite framework for the full account.[16] Similarly, a full account of how conscious thoughts provide reasons for self-ascriptions must make reference to the constitutive connection between the representational content of conscious thoughts, the representational content of self-ascriptions and the content-determining environment. This is too large a project to undertake at this juncture, but again, content externalism provides the requisite framework for the full account.

Before closing this section and turning to the significance of partial representation that lies at the heart of the positive proposal, let us look briefly at the second and third of Peacocke's examples of warranted self-ascriptions. Neither example, according to Peacocke, can be understood as a case of self-ascription warranted by conscious thought. I disagree. Here is the second example, which Peacocke thinks involves, instead, a pre-existing, underlying non-conscious state.

(Case 2) The Case of the Oxford Number

> 'If in some meeting, a practical need emerges to find the phone number of Oxford University, I may think, and/or say, 'I know that the phone number of Oxford University is 270001'. This self-ascription can itself be knowledgeable, but it is not true that it has to be based first on a conscious subjective memory that the number is 270001.' (Peacocke, 1996: 121)

[16] See Burge (2003, 2010). See also Majors & Sawyer (2005).

I agree that in this example the self-ascription is not based on a conscious subjective memory. However, the fact that there is no conscious subjective memory does not mean that there is no conscious state as such that plays a warranting role in the self-ascription. This particular case may in fact be understood as a case where the supposed self-ascription is actually just an expression of belief rather than a self-ascription properly so-called—as a case of endorsement of a propositional content rather than as a report of a psychological state.[17] But the expression of belief is the expression of a conscious state that could warrant an appropriately related self-ascription. More generally, I think self-ascriptions depend on a thought being brought to mind—that is, on the having of a conscious thought. The self-ascription may be based on a pre-existing, underlying non-conscious state, but the self-ascription will be made not directly on that basis, but via the pre-existing, underlying non-conscious state being brought to mind as a conscious thought; and it is the conscious thought that provides the reason for the self-ascription. This second kind of case does not persuade me, then, that there are cases of warranted self-ascription that do not rely on conscious thoughts as their warranting causes.

The third and final kind of case to which Peacocke draws attention involves, he says, neither an intermediate conscious state nor a pre-existing, underlying non-conscious state. These are related to the phenomenon of transparency, and point to cases of self-ascription arrived at through a process of making up one's mind. The example is taken from Evans (1982).

(Case 3) The Case of Transparency

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

[17] This intuitive treatment of the case has affinities with the neo-expressivist views of, for example, Bar-On (2004) and Finkelstein (2008). Despite the affinity in this case, the account of self-knowledge I present in this paper is inconsistent with neo-expressivist views.

While I agree that cases involving transparency are cases of making up one's mind, it doesn't follow from this that there is no conscious thought involved, and hence no conscious thought that could provide the warrant for the relevant self-ascription. If a subject is asked whether she believes *p*, let us suppose that, in the process of making up her mind, she consciously reflects on whether *p*, and hence turns her attention outward to the world rather than inward to her mind. Nonetheless, in coming to a particular view, she acquires a conscious thought that *p*, which she forms after consideration of the facts, and it is this conscious thought that warrants the resulting self-ascription. Thus in these cases too, it is the subject's conscious thought that provides the warrant for her self-ascription.[18]

4. Partial Representation

According to the partial-representation model of self-knowledge, the self-ascription of a thought is authoritative when it is based on, and hence derives its epistemic warrant from, a conscious, occurrent thought in virtue of which it partially represents an underlying state of mind. In the previous section I explained how a conscious thought can provide a reason, and hence a warrant, for a self-ascription. In this section, I explain the significance of partial representation. In the next section, I explain the way in which the notion of partial-representation accommodates behavioural counter-evidence.

In order to understand the nature of partial representation, recall, first, the distinction between thoughts and states of mind. A thought is individuated by an attitudinal relation to a propositional content, where a propositional content is individuated by its constituent concepts. A state of mind, in contrast, is individuated by a thought in conjunction with the way in which, or the extent to which, a subject grasps each of the constituent concepts in the thought's propositional content. A propositional content is therefore a constituent element of a thought, and a thought is a constituent element of a state of mind. Let us say that x *fully represents* y if and only if x represents y as a whole, and that x *partially represents* y if and only if x represents one or more constituent elements of y without fully representing y. Let us also say that x *fully reflects* y if and only if x carries information about y as a whole, and *partially reflects* y if and only if x carries information about one or more constituent elements

[18] For transparency accounts of self-knowledge, see for example Boyle (2009), Byrne (2005, 2011), Fernández (2003, 2013) and Moran (2001). My view is inconsistent with transparency accounts generally, but is consistent with a weak kind of transparency such as that developed in Kind (2003).

of y without fully reflecting y. Using this terminology, let us look, in turn, at first-order thoughts and then self-ascriptions.

The primary function of a first-order thought is to represent a state of the world. The representational constituents of a thought are externally-individuated concepts that relate an individual representationally to a subject matter. As a result, first-order thoughts enable the subject to navigate the world and engage in goal-directed behaviour. But in representing a state of the world, a first-order thought partially reflects an underlying state of mind. It does this by reflecting the subject's attitudinal relation to a propositional content without reflecting the way in which the constituent concepts of the propositional content are grasped. First-order thoughts are directed outwards, towards the world. Self-ascriptions, in contrast, are directed inwards. The primary function of a self-ascription is to represent the subject's state of mind. But a self-ascription is a second-order belief that represents the subject as bearing an attitudinal relation to a propositional content—as having a thought. The way in which she grasps the constituent concepts of the propositional content of that thought is not represented by the representational constituents of the self-ascription. This means that a self-ascription typically represents the subject's state of mind only partially.

One way to think about this is by drawing on the account of warranted self-ascription offered in the previous section. An authoritative self-ascription is based on a conscious thought that functions as an operative reason. An authoritative self-ascription, then, can only represent the information that is carried by the first-order conscious thought on which it is based and from which it derives its epistemic warrant. And if the first-order, conscious thought on which a self-ascription is based only partially reflects the subject's state of mind, the self-ascription itself can only partially represent the subject's state of mind. Insofar as a subject's state of mind involves partial grasp of the constituent concepts in its propositional content, then, an ascription of a state of mind will represent that state of mind only partially.

A second way to think about this is by invoking, once again, a contrastive account of the propositional attitudes. According to a contrastive account of the propositional attitudes, a subject does not think that *p simpliciter*, but thinks that *p* relative to a class of contrasting propositions. The relevant class of contrasting propositions is determined by the way in which, or the extent to which, the subject grasps the constituent concepts of the propositional content of the thought. The propositional content of the thought, however, is itself non-contrastive; what is thought is, simply, that *p*. The contrasting propositions help to individuate the subject's state of mind, but they do not individuate the thought. But since a self-ascription is based on a conscious thought, it will only be able to represent that thought,

and will not be able to represent the contrast class relative to which the subject thinks that *p*. Insofar as a subject's state of mind involves partial grasp of the constituent concepts in its propositional content, then, an ascription of a state of mind will represent that state of mind only partially.

I have said that the self-ascription of a state of mind is authoritative when it is based on, and hence derives its epistemic warrant from, a conscious, occurrent thought in virtue of which it partially represents an underlying state of mind. The claim that a self-ascription can partially represent a state of mind implies the rejection of a simplistic understanding of the distinction between true self-ascriptions on the one hand and false self-ascriptions on the other. We now need to recognise a three-fold distinction between self-ascriptions that are true and complete, self-ascriptions that are true but partial, and self-ascriptions that are false. Self-ascriptions that partially represent a subject's state of mind in the specific sense I have articulated are self-ascriptions that are true but partial.[19] It is the notion of partial representation and the consequent notion of a self-ascription's being true but partial that introduces the possibility of behavioural counter-evidence consistent with first-personal authority. This takes us to the final aspect of the partial-representation account of self-knowledge.

5. Behavioural Evidence and Counter-Evidence

Over the previous two sections, I have argued that a self-ascription is authoritative when it is based on, and hence derives its epistemic warrant from, a conscious, occurrent thought in virtue of which it partially represents an underlying state of mind. In this section, I turn to the role of behavioural evidence and counter-evidence in the partial-representation model of self-knowledge. In particular, I argue that what we ordinarily take to be behavioural counter-evidence to self-ascriptions need not undermine the authoritative, knowledgeable status of such ascriptions when those self-ascriptions are epistemically grounded in conscious occurrent thoughts to which we have privileged access.

There are, as we have seen, two elements to a subject's state of mind. The first element is the thought; the second element is the way in which the subject grasps the constituent concepts in the propositional content of that thought. This allows for the possibility that the epistemic warrant for each element differs. It allows that there are, in

[19] I stick to talk of 'true but partial' self-ascriptions rather than 'partially true' self-ascriptions because the latter way of talking appears to imply that truth comes in degrees. For present purposes, I remain neutral with respect to the question of whether truth comes in degrees.

effect, two ‘epistemic markers’ of a subject’s state of mind. The first epistemic marker is the subject’s conscious thought on which the first-personal self-ascription is based. The second epistemic marker is the subject’s behaviour. The first epistemic marker, the conscious thought, warrants the authoritative self-ascription of a thought. The second epistemic marker, the subject’s behaviour, can provide evidence of the way in which, or the extent to which, the subject grasps the constituent concepts of the propositional content of a thought which she has an authoritative, first-personal reason to self-ascribe. This is because the complex, contrastive nature of a subject’s states of mind, which follows from her incomplete grasp of the concepts involved in her thoughts, manifests itself in her behaviour. As a consequence, her behaviour will provide evidence for her states of mind that are not fully represented by her self-ascriptions.

The partial-representation model of self-knowledge, then, opens up a potential gap between a subject’s state of mind, as evidenced by her behaviour, and her self-ascription of that state of mind, as epistemically grounded in her conscious, occurrent thought. This potential gap in turn opens up the possibility that a self-ascription can be authoritative and yet subject to behavioural counter-evidence. Behavioural counter-evidence is, on standard models of self-knowledge, understood as evidence against the truth of a self-ascription and hence in favour of its falsity. However, this understanding of behavioural counter-evidence is a manifestation of the simplistic understanding of the distinction between true self-ascriptions and false self-ascriptions that I rejected earlier. In its place, we now have a three-fold distinction between self-ascriptions that are true and complete, self-ascriptions that are true and partial, and self-ascriptions that are false. This inevitably alters the way in which we should understand behavioural counter-evidence. Behavioural counter-evidence is, on the partial-representation model of self-knowledge, evidence against a subject’s self-ascription being both true and complete. But evidence against a subject’s self-ascription being both true and complete is not yet evidence against the truth of that self-ascription, and hence not yet evidence in favour of its falsity. This is because the fact that a self-ascription is not both true and complete is consistent with its being false, but is also consistent with its being true and partial.

Whether a pattern of behaviour indicates the true but partial nature of a self-ascription or its outright falsity will depend on the extent and breadth of conflict involved. In cases where the subject’s behaviour is inconsistent, sometimes supporting the self-ascription and sometimes going against it, it is natural to assume that the relevant self-ascription is true but partial. In cases where the subject’s behaviour goes consistently against the self-ascription, it

is natural to assume that the relevant self-ascription is simply false. Although I will not argue the point here, cases of the latter kind are, I think, likely to involve systematic self-deception or pathology. Cases of the former kind, in contrast, are relatively commonplace, and it is cases of this kind that can help to illustrate the theoretical benefits of the partial-representation model of self-knowledge. Consider the following example from Schwitzgebel.[20]

Schwitzgebel asks us to imagine Ralph, a Philosophy Professor, who sincerely professes to believe that men and women are equally intelligent, arguing coherently, consistently and authentically for the view, despite the fact that he is prone to sexism in his behaviour.[21] For example, he tends to think that his male students are more insightful than his female students, is more surprised if a female colleague offers an interesting comment in discussion than if a male colleague does, and requires more evidence to be convinced that a woman is capable of filling a position in the department than that a man is. Schwitzgebel says:

> Ralph's attitude toward the intellectual equality of the sexes is what I would call an in-between state. His dispositions, his patterns of response, his habits of thought, are mixed up and inconsistent. It is neither quite right to say that he believes in the intellectual equality of the sexes nor quite right to say that he fails to believe that. But he has no specially privileged self-knowledge of the fact. (Schwitzgebel, 2012: 192).

I agree with Schwitzgebel that Ralph's attitude toward the intellectual equality of the sexes is 'an in-between state' and that it is 'neither quite right to say that he believes in the intellectual equality of the sexes nor quite right to say that he fails to believe that'. This is to be explained, on the account I am suggesting, by the fact that Ralph is in a complex, contrastive state of mind. Ralph believes that the sexes are intellectually equal relative to some contrasting propositions, but not relative to other contrasting propositions. Thus adopting a contrastive account of the propositional attitudes provides a concrete framework within which we can understand such 'in-between' states. I also agree with Schwitzgebel's claim that Ralph has 'no specially privileged self-knowledge' of the fact that he is in such an in-between state. This is to be explained, on the account I am suggesting, by the fact that a

[20] See Schwitzgebel (2012).

[21] The example can be understood as a case of implicit bias, which, as a result, I do not treat separately.

subject's conscious, occurrent thoughts, which, when operative, provide reasons for her self-ascriptions, reflect her underlying states of mind only partially. What is attributed in a self-ascription is the subject's attitude to a non-contrastive proposition, but her underlying state of mind is essentially contrastive. Self-ascriptions, then, cannot provide a subject with privileged or authoritative knowledge of her contrastive states of mind.

But I reject Schwitzgebel's claim that examples of this kind demonstrate our self-ignorance in the sense of undermining privileged or authoritative self-knowledge *per se*.[22] Ralph, despite the inconsistencies in his dispositions, patterns of response and habits of thought, knows that he believes in the equality of the sexes. This can seem counterintuitive. But its counterintuitive appearance is due to the assumption, embedded within and perpetuated by traditional theories of self-knowledge, that knowledgeable self-ascriptions are both true and complete. This assumption is in part what drives the intuition that in order to be authoritative, self-knowledge must be immune to behavioural counter-evidence. Although the intuition of immunity to behavioural counter-evidence is no longer widespread, the assumption that drives it remains deeply embedded. The partial-representation model of self-knowledge I have been advocating rejects not only the intuition but the driving assumption. According to the partial-representation model of self-knowledge, a self-ascription can (and typically does) occupy the 'middle ground' of being true and partial. It follows that Ralph can really know that he believes in the equality of the sexes, even though his self-ascription does not provide a true and complete representation of his underlying state of mind. The introduction of the notion of partial representation thus allows us to make sense of the knowledgeable status of self-ascriptions in the light of behavioural counter-evidence. On the account I have offered, the authority of a self-ascription derives from its being epistemically grounded in a conscious thought to which the subject has privileged access, and this authority is consistent with the self-ascription being true but partial, and hence consistent with behavioural counter-evidence. It is in this sense that the partial-representation model of self knowledge offers a middle path between the two extremes of absolute epistemic security on the one hand and scepticism about first-personal self-knowledge on the other.

6. Conclusion

[22] Schwitzgebel offers a number of different examples that supposedly demonstrate our self-ignorance. I do not address the array of examples here, but think the partial-representation model of self-knowledge has the resources to explain them all. This is a project for another occasion.

In this paper, I have offered an account of self-knowledge according to which the self-ascription of a thought is authoritative when it is based on, and hence derives its epistemic warrant from, a conscious, occurrent thought in virtue of which it partially represents an underlying state of mind. Whether or not the partial-representation model of self-knowledge is accepted, it does have two virtues. First, it accommodates the two distinctions that follow from content externalism: that between concepts and conceptions, and that between thoughts and states of mind. Second, it offers a principled way to navigate a middle path between two extremes: that of absolute epistemic security on the one hand, and that of scepticism about first-personal self-knowledge on the other. Whether other theories of self-knowledge can be adapted to do so remains to be seen.

References

Bar-On, D. (2004) *Speaking My Mind: Expression and Self-Knowledge* (Oxford: Oxford University Press).

Boyle, M. (2009) 'Two Kinds of Self-Knowledge,' *Philosophy and Phenomenological Research* 78: 133–64.

Burge, T. (1979) 'Individualism and the Mental,' in P. French, T. Uehling & H. Wettstein (eds) *Midwest Studies in Philosophy* 4 (Minnesota: Minnesota University Press).

Burge, T. (1986) 'Intellectual Norms and Foundations of Mind,' *Journal of Philosophy* 83: 697-720.

Burge, T. (1993) 'Content Preservation', *Philosophical Review* 103: 457-88.

Burge, T. (2003) 'Perceptual Entitlement,' *Philosophy and Phenomenological Research* 67: 503-48.

Burge, T. (2010) *Origins of Objectivity* (Oxford: Oxford University Press).

Burge, T (2013) 'Epistemic Warrant: Humans and Computers', in his *Cognition Through Understanding* (Oxford: Oxford University Press).

Byrne, A. (2005) 'Introspection,' *Philosophical Topics* 33: 79–104.

Byrne, A. (2011) 'Transparency, Belief, Intention,' *Aristotelian Society Supplementary Volume* 85: 201–21.

Cassam, Q. (2014) *Self-Knowledge for Humans* (Oxford: Oxford University Press).

Descartes, R. (1641) *Meditations on First Philosophy*, translated by John Cottingham (Cambridge: Cambridge University Press, 1996).

Dretske, F. (2000) ‘Entitlements: Epistemic Rights Without Epistemic Duties?’ *Philosophy and Phenomenological Research* 60: 591-606.

Evans, G. (1982) *The Varieties of Reference* (Oxford: Oxford University Press).

Fernández, J. (2003) ‘Privileged Access Naturalized,’ *The Philosophical Quarterly* 53: 352–72.

Fernández, J. (2013) *Transparent Minds: A Study of Self-Knowledge* (Oxford: Oxford University Press).

Finkelstein, D. (2008) *Expression and the Inner* (Cambridge, MA: Harvard University Press).

Goldberg, S. (2015) *Externalism, Self-Knowledge and Scepticism* (Cambridge: Cambridge University Press).

Kind, A. (2003) ‘What’s so Transparent about Transparency?’ *Philosophical Studies* 115: 225-44.

Kripke, S. (1972) ‘Naming and Necessity,’ in Donald Davidson & Gilbert Harman (eds) *Semantics of Natural Language* (Dordrecht: Reidel).

Kripke, S. (1980) *Naming and Necessity* (Cambridge MA: Harvard University Press).

Ludlow, P. & Martin, N. (1998) *Self-Knowledge and Externalism* (Chicago: CSLI Publications).

Majors, B. & Sawyer, S. (2005) ‘The Epistemological Argument for Content Externalism,’ *Philosophical Perspectives* 19: 257-280.

McDowell, J. (1994) *Mind and World* (Cambridge MA: Harvard University Press).

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

Peacocke, C. (1996) ‘Entitlement, Self-Knowledge and Conceptual Redeployment,’ *Proceedings of the Aristotelian Society* 96: 117-58.

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

Pessin, A. & Goldberg, S. (1996) *The Twin Earth Chronicles* (New York: M.E. Sharpe).

Plantinga, A. (1993) *Warrant: The Current Debate* (Oxford: Oxford University Press).

Putnam, H. (1973) ‘Meaning and Reference,’ *The Journal of Philosophy* 70: 699-711.

Putnam, H. (1975) ‘The Meaning of “Meaning” ,’ *Minnesota Studies in the Philosophy of Science* 7: 131-93.

Ryle, G. (1949) *The Concept of Mind* (Chicago: University of Chicago Press).

Sawyer, S (2014) ‘Contrastive Self-Knowledge’, *Social Epistemology* 28: 139-152.

Sawyer, S (2015) ‘Contrastive Self-Knowledge and the McKinsey Paradox’, in S. Goldberg (ed.) *Externalism, Self-Knowledge and Skepticism* (Cambridge: Cambridge University Press): 75-93.

Sawyer, S. (2018) ‘The Importance of Concepts’, *Proceedings of the Aristotelian Society,* 118: 127-47.

Sawyer, S. (2019) ‘Talk and Thought’, in Alexis Burgess, Herman Cappelen & David Plunkett (eds) *Conceptual Engineering and Conceptual Ethics* (Oxford: Oxford University Press), forthcoming.

Schwitzgebel, E. (2012) ‘Self-Ignorance’, in JeeLoo Liu & John Perry (eds) *Consciousness and the Self* (Cambridge University Press).

Shoemaker, S. (1988) ‘On Knowing One’s Own Mind’, in James E. Tomberlin (ed.) *Philosophical Perspectives, 2, Epistemology* (Atascadero, Cal: Ridgeview Publishing Company).