Franz Knappik

Universitetet i Bergen

Sellars on Self-Knowledge

Wilfrid Sellars had an elaborate theory of self-knowledge about one's own thoughts that anticipates some crucial claims and topics of current work on self-knowledge. In this contribution, I reconstruct Sellars's theory of self-knowledge, and explore connections with more recent work on the topic. I argue that Sellars's account undermines Shoemaker's and Burge's influential arguments against "perceptual" accounts of self-knowledge, and I discuss whether Sellars's position is apt to give a plausible account of the relation between self-knowledge and phenomenal consciousness.

In the last sections of *Empiricism and the Philosophy of Mind*, Wilfrid Sellars uses the famous Myth of Jones in order to present a novel framework for thinking about the mind. Central to Sellars's proposal is an original account of the self-knowledge that we have regarding our own thoughts and sense impressions. In a penetrating and thought-provoking correspondence with Hector-Neri Castañeda, Sellars further develops his remarks on self-knowledge about thoughts in EPM into a rich and complex account. This account anticipates in many ways ideas and topics in the more recent literature on self-knowledge. It is my aim in this paper to reconstruct Sellars's views on self-knowledge, and to explore some of its relations to the contemporary debate.

When discussing self-knowledge of one's own present thoughts, Sellars builds on our intuitive notion of a "thought" that stands for occurrent mental states with conceptual content (as opposed to the dispositional attitudes that we can take towards such contents, for instance beliefs). Like other mental states and episodes, thoughts are normally considered as being accessible to their subjects in a particular way, which differs from the access we have to other

persons' thoughts. Thus, we seem to have an *immediate* or non-inferential knowledge of our own thoughts—for instance, we do not have to observe and interpret our own behavior, facial expressions etc. in order to know our thoughts. Moreover, we seem to be in a particularly good *epistemic position* with regard to our own thoughts. We would expect that our beliefs about what we are currently thinking are very likely true, while our beliefs about other persons' thoughts may easily be mistaken.[1] Sellars shares the view that ordinary self-knowledge about thoughts is special in these ways. Indeed, his entire account of self-knowledge is motivated by the ambition to do justice to our intuitions regarding those special features, while at the same time avoiding the Myth of the Given that, on Sellars's view, traditional accounts of privileged access have fallen victim to.

The paper falls into three parts. In section 1, I will reconstruct Sellars's theory of self-knowledge about thoughts. Section 2 examines how Sellars's discussion bears on Shoemaker's and Burge's influential arguments against "perceptual" accounts of self-knowledge. In section 3, I contrast Sellars's view with recent work that assigns phenomenal consciousness an important role in the explanation of self-knowledge about thoughts, and I propose to amend his account by combining it with a Higher-Order Thought theory of phenomenal consciousness.

1. Sellars's Account of Self-Knowledge of One's Thoughts[2]

In the last part of EPM, Sellars uses the famous Myth of Jones in order to identify a middle ground between two extremes which he finds equally unattractive. The one extreme is occupied by classical views of the mind in both the rationalist and the empiricist traditions. According to such views, there really are inner episodes and states such as thoughts, perceptions, beliefs and intentions. All or at least some of these inner phenomena are the

[1] Cf., e.g., Shoemaker 1994; Burge 1996; Moran 2001. For skepticism, see Carruthers 2011. For a detailed overview and assessment of the debate, see Gertler 2010.

[2] Johannes Haag offers an excellent discussion of Sellars's account of self-knowledge, from which I have greatly profited, in Haag 2001. I signal some points of disagreement with Haag in footnotes below.

object, on the traditional view, of a knowledge with special and perhaps even infallible authority, which can serve as a foundation for all further knowledge. Such views are not acceptable to Sellars, who has argued in the previous parts of EPM that they share a commitment to the Myth of the Given. Arguably, the basic claim of the Myth of the Given is that there can be entities that provide epistemic support to rational attitudes without being themselves epistemically dependent on anything else (deVries and Triplett 2000, xxvi). It is easy to see that the traditional assumption of a kind of self-knowledge that can serve a foundational role fulfils these criteria: on this assumption, the mere occurrence of certain mental states can justify beliefs *about* those states (which then can justify further beliefs), while the justificatory force of those states is not itself derived from any further factors.

Opposed to such views of self-knowledge is the equally radical view of behaviorism. Behaviorism denies that inner states and episodes have any existence in their own right—rather, they are, on this view, reducible to behavioral episodes and dispositions. As a consequence, there can be no substantial epistemological difference between first- and third-personal access to inner states and episodes. Our knowledge about them is always inferred on the basis of observable behavioral evidence, and the first-personal case differs from the third-personal case only in that we normally have more such evidence available about ourselves than about other persons.[3] Sellars rejects both the ontological and the epistemological claims of behaviorism: he wants to defend both the irreducible existence of inner states and episodes, and the idea that our first-personal access to the inner is special in some ways. In particular, Sellars thinks that we have indeed a non-inferential, epistemically very good and distinctively first-personal access to our own thoughts and sense impressions.

The last sections of EPM therefore face the task of showing how one can accept such a special first-personal access to really existing inner episodes *without* falling victim to the Myth of the Given. Sellars's response consists of three major moves. The first move is to

[3] The classical statement of this position is Ryle 1949, ch. 6.

interpret the vocabulary of inner episodes such as thoughts and sense impressions as being originally a *theoretical* vocabulary. Thoughts, in particular, are postulated in the Myth of Jones as theoretical entities that are meant to explain episodes of overt intelligent behavior. As Sellars is eager to emphasize, he does not understand the distinction between theoretical and observable entities as an *ontological* distinction: by classifying something as a theoretical entity, one does not imply that it does not really exist (EPM §43: 173–74). Rather, one assigns it a different place in one's theories. Unlike an observable entity, a theoretical entity is not "definable in observational terms" (EPM §58: 187), but this neither entails that it is not part of reality, nor that there is not sufficient reason to believe in its existence (EPM §58: 187).

As is often the case with theoretical concepts, Jones's concept of a thought is formed by analogy with observational concepts—in the case of thoughts, concepts which designate episodes of *overt speech* (EPM §56: 186). Hence, according to Jones's theory, all intelligent overt behavior is ultimately caused by inner episodes that are analogous to episodes of overt speech (EPM §56: 186).[4]

Sellars's second move in accounting for first-personal access is to show how a first-personal use of Jones's theoretical concept of a thought can come about. After Jones has taught his Rylean fellows how to explain each other's behavior in terms of their thoughts, he trains them to apply the concept of a thought to themselves. First, the trainee—whom Sellars calls Dick—learns to use his own behavior as evidence for self-ascriptions of thoughts: Dick is conditioned to respond to such evidence in the same way as an observer would do. This yields a form of access to one's own thoughts which does *not* yet display the special features of the first-personal perspective. It is not immediate, as Dick has to draw an explanatory

[4] Sellars's idea that concepts like that of a thought are originally theoretical concepts has been enormously influential in the philosophy of mind and in psychology. Sellars is often cited as the first author to have proposed a so-called "Theory Theory" of mindreading (e.g. Lewis 1972, 257; Nichols and Stich 2003, 7). Mindreading is the ability to interpret, explain and predict the behavior of other creatures in terms of mental states, and the Theory Theory of mindreading holds that this ability should be understood as a typically implicit, folk-psychological *theory*. Cf. O'Shea 2012 for discussion of the relation between Sellars's views and the Theory Theory of mindreading.

inference from his own observed behavior to a thought that he postulates as an explanation for his behavior. And this form of access does not provide a particularly good epistemic situation, as an external observer can have exactly the same access to Dick's thoughts—indeed, the conditioning that takes place here is precisely a process in which Dick is supposed to take over Jones's essentially third-personal access to his thoughts.

Nevertheless, Sellars thinks that if there is appropriate training, a new situation can arise which *does* include a distinctively first-personal access. In EPM, Sellars characterizes this new situation only very briefly as follows:

> And it now turns out—need it have?—that Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behavior. Jones brings this about, roughly, by applauding utterances by Dick of 'I am thinking that p' when the behavioral evidence strongly supports the theoretical statement 'I am thinking that p'; and by frowning on utterances of 'I am thinking that p', when the evidence does not support this theoretical statement. (EPM §59: 189)

If this training succeeds, Sellars points out, the language that first had a purely theoretical use "has gained a reporting role", and our ancestors "begin to speak of the privileged access each of us has to his own thoughts" (EPM §59: 189).

However, the brief discussion of this topic in EPM fails to make it clear how this result can precisely be brought about, and what exactly sets apart the training that Sellars describes here from the first stage of Dick's training, where he was conditioned to self-interpret his own behavior. Luckily for us, Hector-Neri Castañeda has challenged Sellars in an extensive correspondence to be more explicit about the details of his account of self-

knowledge.[5] As becomes clearer from that correspondence, Sellars assumes that privileged access comes about if the process of training establishes a causal connection between a subject's thoughts and his self-ascriptions which *does not* involve perception of the behavior that had previously served as evidence for the self-ascriptions. Thus, when discussing Castañeda's analogy of a process in which Dick learns to report the presence of a colony of viruses in his kidney that causes determinate symptoms, he describes the following scenario:

> [1.] We ask Dick to say 'I am in state φ' [i.e., in a state in which the viruses are present] every two minutes or so.
>
> [2.] Sometimes Dick is in state φ, sometimes he isn't.
>
> [3.] When he is, *we* know that he is by observing symptoms U_i . . . U_j and using a well confirmed theory.
>
> [4.] We put him in a position where he cannot observe these symptoms.
>
> [5.] When he says 'I am in state φ' and actually is in state φ we reward him; in the contrary case we give him a mild shock.
>
> [6.] It turns out that for some states φ (but by no means for all) we can bring about a connection between being in state φ and saying 'I am in state φ'. (To Castañeda, 14.11.1969, §4)

As this passage shows, Sellars thinks that in the training that brings about *immediate* self-ascriptions, the trainee has to be precluded from perceptual access to the behavioral symptoms that warrant the trainer's ascriptions of mental states. As a consequence, the training, if successful, will establish a direct causal link between the trainee's first-order thoughts and his self-ascriptions which is not mediated by any observation of his own behavior.

[5] Sellars and Castañeda 2006. In references to the Sellars-Castañeda correspondence, I cite dates of letters and paragraph numbers.

If such a causal link can actually be established, Dick thereby acquires a novel way of self-ascribing his thoughts, which differs from the earlier, evidence-based way in two important respects. First, it is non-inferential: Dick can now self-ascribe his thoughts without having to draw explanatory inferences on the basis of observed behavior. Second, it is specifically first-personal insofar as it is based on a causal mechanism which (at least in worlds that are sufficiently similar to ours) can obtain only among the thoughts of the *same* person, not among different persons (to Castañeda, 3.4.1961, §15).

Once the subject has acquired this new way of making *overt* self-ascriptions of thoughts, it is easy to see how a subject can also acquire an ability for reliable purely mental self-ascriptions, or "meta-thoughts", as Sellars calls them. For according to Jones's theory, a statement like "I am thinking that p" has as its immediate cause a thought with the same content. So the mechanism that is established by the conditioning in question will involve the causation of meta-thoughts anyway, and where the subject suppresses the overt self-ascription, a purely mental self-ascription will result.

What remains to be done in the third part of Sellars's account is to explain how precisely our self-knowledge is *justified*, or how our meta-thoughts acquire *epistemic authority*, and in particular how they can acquire the particularly *strong* authority that seems to characterize normal self-knowledge. While he is silent about this part of his theory of self-knowledge in EPM, he develops it in some detail in the correspondence with Castañeda.[6] As Sellars points out, "the key to the account I have given [in EPM] of the direct-noninferential knowledge of inner episodes is the apparatus I developed in discussing the status of noninferential perceptual knowledge" (to Castañeda, 3.4.1961, §13). Put very briefly, Sellars thinks that non-inferential perceptual knowledge has an epistemic authority that is depends on a background of assumptions and conceptual abilities. For instance, the epistemic authority attached to normal perceptual reports like "Lo! Here is a red apple!" presupposes a

[6] For the following, cf. Haag 2001, 297–98, 312–3.

background which includes knowledge that such reports are *reliable indicators* of the actual presence of red apples under standard perceptual conditions, as well as the knowledge that currently, such standard conditions obtain. It is this background which enables competent subjects to draw so-called "trans-level inferences" like the inference from

(Perc-1) I have uttered "Lo! Here is a red apple!" under standard conditions of perception

to

(Perc-2) There is good reason to believe that there is a red apple in front of me.

According to Sellars, the *availability* of such trans-level inferences—a subject's *ability* to draw them—justifies the epistemic authority of the normal perceptual reports and beliefs.

How can the epistemology of self-knowledge be explained along parallel lines? Sellars describes the trans-level inference for the case of self-knowledge as inference from a premise of the form

(1) The thought that I am thinking that p has occurred to me in such and such a manner and context

to a conclusion of the form

(2) I am thinking that p

(to Castañeda, 8.12.1961, Ad II.2).[7]

In a related context, Sellars clarifies two important points. First, Sellars explains that the meta-thought

(MT-1) I am thinking that it is raining

[7] Haag 2001, 312 argues that in order for the inference to be a trans-level inference, the conclusion should read "There is good reason to believe that I am thinking that p". But Sellars's formulation of the inference as it stands matches the way he introduces the term "trans-level inference" at PH 88.— Sellars oscillates in his characterizations of the resulting self-knowledge between present-directed formulations like (2), and formulations directed to the immediate past (e.g. to Castañeda, 3.4.1961, §17: "I (just) had the thought that-p"). I will bracket this difference in my discussion, and always talk about present-directed self-knowledge.

is "self-referentially incorrigible", i.e., it is automatically true insofar as the thought "It is raining" is itself a part of this meta-thought, and is thought by the subject whenever she thinks the meta-thought.[8] The same holds, mutatis mutandis, for the third-order thought

(MT-2) I am thinking that I am thinking that it is raining.

It seems that Sellars wants to set *this* kind of trivial self-knowledge aside, and *therefore* prefers the above formulation (1) to a formulation like (MT-2) as means to self-ascribe a meta-thought.[9] (Of course, Sellars must understand the phrase "has occurred" as referring to a period that started a very short time ago. One might say instead "has *just* occurred" to highlight this.)

Second, Sellars seems to assume that the subject needs to specify a particular "manner and context" in premise (1) only insofar as she needs to rule out conditions of massively impaired rationality, in which an episode that *seems* to be a thought that p is actually *not* a thought, but merely "mental noise" (to Castañeda, 11.3.1962, Ad 3; Ad 6).[10] We therefore can replace (1) by the conjunction of the following two premises:

(1a) The thought that I am thinking that p has occurred to me

(1b) My rationality is not currently massively impaired.

Moreover, Sellars seems to assume that knowledge about (1b) is easily had. That our rationality is not presently massively impaired is an assumption that is built in into every proper thought: for once again, if rationality *would* be massively impaired, there would be no real thoughts, but only "mental noise".

[8] Sellars anticipates here Burge's idea of "cogito-like thoughts" (Burge 1996, 92).

[9] Elsewhere, Sellars works with the formulation "I am under the impression that I am thinking that p" (to Castañeda, 11.3.1962, Ad 1).—Sellars distinguishes in this context also between mere thoughts and "mental assertions" (cf. Haag 2001, 308–11), where a thought with the content "It is raining" is a mental assertion of "It is raining" "unless it occurs in a larger context (e.g. 'It is raining or snowing') which suspends its assertive force" (to Castañeda, 3.3.1962, Ad 3c). Sellars implies that a meta-thought like "I am mentally asserting that p" is not self-reflectively incorrigible. However, when such meta-thoughts are read in the sense "I am *hereby* mentally asserting that p", they *do* seem to display self-referential incorrigibility. And conversely, meta-thoughts about *mere* thoughts can lack self-referential incorrigibility if they have the right kind of content—as in (1) and (1a) above. I therefore bracket the issue of mental assertion in the following.

[10] Sellars tends to formulate this condition as self-ascription of a "well-functioning" rationality (e.g. to Castañeda, 11.3.1962, ad 3.c), but cf. his reference to the "more extreme forms" of "madness" (ibid.) in the same context.

Now if a subject S has undergone the conditioning described above, it will actually be the case that

(3) S is so conditioned that, if her rationality is not massively impaired, the thought that she is thinking that p occurs to her (ceteris paribus) only when she is actually thinking that p.

But the truth of (3) does not on its own suffice in order for S to be entitled to make the above trans-level inference from (1a) and (1b) to (2). Rather, it is necessary that S also *know* that she is entitled to make this inference, or that (3) is true. For as in the case of perceptual justification, Sellars thinks that a subject's belief that (2) is justified only if the subject is *able* to make the trans-level inference from (1a) and (1b) to (2), and this ability requires a knowledge about the legitimacy of such an inference (cf. EPM §35: 168).

So how can a subject know that (3) is true? Sellars does not explicitly address the question, but his exchange with Castañeda does contain the resources for a possible answer. This answer builds on the assumption of an essential connection between *rationality* and the existence of a reliable causal link between thoughts and meta-thoughts. As Sellars writes:

> [N]on-inferential autobiographical knowledge of what one is thinking . . . is essential to rationality. In short, the ability to think in the full sense of the term involves the ability to meta-think, and the ability to have noninferential knowledge of what one is thinking. (to Castañeda, 8.12.1961, Ad II.3)

Why should it not be possible for a creature to have robust capacities for rational thought and intelligent behavior without having the resources to form meta-thoughts, and the ability for non-inferential knowledge of its thoughts? The idea that there is such a connection seems to derive from Sellars's more general views on the nature of language use and epistemic activity. For one thing, Sellars holds that one becomes a "full-fledged member of the linguistic

community" (MFC 86) only once one is able to think meta-thoughts about linguistic items, for only then is one able to teach one's language to others (ibid.). In addition, Sellars famously claims that empirical knowledge is a "self-correcting enterprise" (EPM §38: 170). He seems to assume that someone can self-correct, if necessary, his judgments and inferences only if he is able to know in the first place what his judgments and inferences *are*; and he holds that *inferential* self-knowledge would not suffice for this purpose. While Sellars does not explicitly argue for this latter point, his motivation for it is presumably that inferential self-knowledge of thoughts depends on appropriate evidence for the presence of a thought, that such evidence is not always available, and that we need to be able to exercise rational control also over all the thoughts for whose presence we have no sufficient evidence. Taken together, these points support the idea that rationality requires knowledge that inferences of the above form (from (1a) and (1b) to (2)) are normally legitimate (or that a general proposition like (3) holds). We therefore need to acquire such knowledge as part of the broader web of conceptual and inferential abilities that we are taught in the process of becoming rational beings. As Sellars puts it with respect to the conditioning process described above: "the conditioned person comes to conceive of himself as so conditioned" (to Castañeda, 11.3.1962, Ad 6).[11]

If this is how Sellars understands the justification of our self-knowledge about our thoughts, the following question arises: does not the warrant that we get for second-order thoughts or beliefs with the content "I am thinking that p" through such trans-level inferences itself presuppose knowledge at an even higher order—namely, knowledge that a premise like (1a) is true? And if so, doesn't Sellars's account of self-knowledge fall victim to a vicious regress? Castañeda raises this worry in his letter of January 17th, 1962 (§10). To this, Sellars replies that he has never claimed that one must actually *draw* a trans-level inference in order

[11] Notice that according to Sellars, humans are already rational in the pre-Jonesean age, and therefore have already non-inferential self-knowledge of their utterances and their dispositions for utterances at that time: to Castañeda, 8.12.1961, Ad II.3; cf. Haag 2001, 326.

to be justified in believing its conclusion—rather, one only must be *able* to draw it: "I know a vicious regress when I see it" (to Castañeda, 11.3.1962, Ad 10).

Spelt out somewhat more explicitly, Sellars's reply to the regress objection can be understood as follows. Our entitlement to self-knowledge requires indeed that, on a given occasion, we have knowledge about the first premise in the trans-level inference whose availability justifies our self-knowledge. But in order for our belief in the first premise to be justified, it is enough that we are able to draw a further trans-level inference which has that premise as its conclusion. Thus, if challenged, we need to be able to justify the belief "The thought that I am thinking that p has occurred to me" through a further trans-level inference. Of course, the first premise of *this* further argument could be challenged, too, and the subject has to be able to defend *this* premise, too, by yet another argument of the same form. But since the availability of those additional arguments is sufficient in order for the subject to be justified in believing their conclusions—and ultimately, in believing the original conclusion "I am thinking that p"—no infinite regress of actual inferences arises. (Of course, Sellars is relying here on his anti-foundationalist view that we "can put *any* claim in jeopardy, though not *all* at once" (EPM §38: 170).)

Notice, however, that this response to the regress objection works only on a particular understanding of what it means that a subject is *able* to draw some inference. On *one* natural understanding of this phrase, that a subject is able to draw an inference requires that the subject *know* the premises of the inference. This cannot be the understanding that Sellars has in mind. For in that case, the subject would be able to justify the second-order premise

(1a) The thought that I am thinking that p has occurred to me

through a trans-level inference only if she already knew, and hence also *believed*, the third-level content

(4) The thought has occurred to me that the thought has occurred to me that I am thinking that p.

The same would apply to the justification for the subject's knowledge about this latter, third-level content: that the subject be *able* to draw yet a further trans-level inference in support of this third-level belief would require her to possess a relevant fourth-level belief, and so on. Justification for self-knowledge about the first-order thought "p" would therefore require an infinite number of higher-order beliefs, and hence be psychologically impossible for finite minds. This consequence can only be avoided if the subject's *ability* to draw the requisite trans-level inferences is understood in another way, namely such that it is enough for this ability if the subject is *in a position* to know the relevant premises. On this view, when I believe that

(2) I am thinking that p,

I have justification for this belief if I am in a position to know (1a), and am able to use this as a premise in a trans-level inference of the above form. In order to justify my belief in (1a) when challenged, I need to be in a position to know that (4), and be able to use this as a premise in a relevant trans-level inference, and so on. (This presupposes that lower-order thoughts normally trigger corresponding higher-level thoughts whenever a question about the existence of the lower-order thought arises; but it is natural to assume anyway that this is a feature of the causal mechanism that links thoughts to meta-thoughts.[12])

So this is how I would reconstruct Sellars's account of self-knowledge about our thoughts. There are several points in Sellars's argument that can be challenged quite easily, but these I will mention here only to put them aside. First, the argument rests on a strong claim about the relation between rationality and self-knowledge. Against the idea that rationality requires self-knowledge, it may be argued that substantive forms of rationality can be realized in the absence of an ability for self-knowledge (cf. Peters 2014, 149 on autism).

[12] At one point of the correspondence with Castañeda, Sellars even suggests that "[o]ne is conditioned to react by having MΘ [i.e. meta-thoughts] *when* (ceteris paribus) and only when one has Θ [i.e. the corresponding first-order thought]" (to Castañeda, 11.3.1962, Ad 1, emphasis added).

Nevertheless, claims like Sellars's have been made quite frequently in the contemporary debate (e.g. Shoemaker 1994; Burge 1997; Moran 2001), and I will not challenge Sellars's position on this front in what follows.

Furthermore, many philosophers and psychologists today would hold that causal mechanisms like the one that supports non-inferential self-knowledge on Sellars's account may be innate as the result of evolutionary selection (e.g. Nichols and Stich 2003, 162, about a very similar mechanism). If this view is correct, there is no need to speculate about what kind of training may bring about such mechanisms.

Finally, it is natural to object to Sellars's account that it is overly demanding. In particular, do rationality and self-knowledge really require that one has background knowledge about the functioning of one's rationality (premise (1b)), and hence also that one possesses a concept of rationality? It may be possible to weaken Sellars's account in this respect: even by Sellars's lights, it may suffice for the justification of self-knowledge that (1b) is true even though the subject doesn't know this. But once again, further discussion of this point would go beyond the limits of this article.

2. Sellars and Arguments against Perceptual Accounts of Self-Knowledge

As we have seen, Sellars ascribes a central role in his account of self-knowledge to a causal mechanism that connects our first-order thoughts and our corresponding meta-thoughts, and that we acquire through appropriate training. Hence, Sellars holds a version of what Shoemaker 1994 calls a "broad perceptual model" of self-knowledge (cf. Haag 2001, 319–24): a view that treats self-knowledge as analogous to perceptual knowledge insofar as it understands it as being based on causal link between a representation and its object. In the last twenty years or so, many authors have criticized perceptual views of self-knowledge (about

thoughts, but also about other mental states), and proposed alternative, non-perceptual views instead.[13]

Probably the most influential arguments against perceptual accounts are the objections that have been raised by Sydney Shoemaker and Tyler Burge (Shoemaker 1994; Burge 1996). Both authors have argued that any adequate explanation of self-knowledge has to take into account the fact that there is a close connection between self-knowledge and *rationality*. The connection that they have in mind is very similar to the one we have encountered in Sellars's discussion (although neither Shoemaker nor Burge seem to be aware of Sellars's remarks on this connection): rationality requires the capacity to engage in self-critical reasoning, to control one's own thoughts and attitudes in the light of one's reasons. Yet such self-critical reasoning is possible only for a creature that is able to know what thoughts it is thinking, and what attitudes it has. Now this connection between rationality and self-knowledge, Burge and Shoemaker argue, is incompatible with a view on which self-knowledge is based on a causal link between first- and second-order states or episodes. On Shoemaker's version of the argument (Shoemaker 1994), such a view has to allow for the possibility that the causal link in question breaks down (resulting in a condition of "self-blindness"), so that the subject is rational, but not anymore capable of self-knowledge. Yet this implication is incompatible with the connection that obtains between rationality and self-knowledge, so the assumption of self-knowledge being based on a causal link must be resisted.

Burge offers a related argument (Burge 1996). If the link between first- and second-order state were only causal, he argues, both could come apart, and we would have to expect the possibility that our second-order thoughts are mistaken without a fault of our own—the possibility of what Burge calls a "brute error". In that case, however, detecting through one's second-order judgment a shortcoming in one's first-order states or episodes would not immediately give one reason to revise the first-order state or episode: the possibility of a brute

[13] For an overview, see Gertler 2010, chs. 6 and 8. (Gertler calls non-perceptual views "rationalist", and perceptual views "empiricist".)

error would always offer a possible excuse that releases one from one's obligation to revise the first-order state or attitude (Burge 1996, 109–10). Yet it is precisely central to self-knowledge in the context of critical reasoning, Burge argues, that the contents of that self-knowledge *can* give us immediate reason to revise our first-order states or attitudes. The ability to detect and hence amend shortcomings at the first-order level is exactly what self-knowledge must contribute to rationality. Therefore, Burge concludes, a non-causal, conceptual connection has to obtain between knowledge and its objects in the case of rational self-knowledge.

Strictly speaking, Shoemaker and Burge do not exclude that the first-order and the second-order state are *also* connected by a causal mechanism. What their arguments are meant to show is that this mechanism cannot play any essential role in the epistemology of self-knowledge. Indeed, Shoemaker's and Burge's arguments have convinced many authors that this conclusion is correct. In the wake of their work, various alternative proposals for how else the connection between first- and second-order state could be understood have been suggested. Several authors, including Shoemaker himself, have argued that a second-order awareness or self-ascription forms a *constitutive* part of the first-order state: it is of the essence of the first-order state, on this account, to be accompanied by a self-ascribing higher-order state (Shoemaker 1994, 288). In the next section I will mention further views which are meant to offer alternatives to a perceptual account of self-knowledge of one's thoughts.

I believe that one central lesson that the contemporary debate on self-knowledge can draw from Sellars is that the opposition between perceptual and non-perceptual models of self-knowledge is not well motivated, and that Shoemaker's and Burge's influential arguments should be seen with more skepticism. For on the one hand, Sellars agrees with Burge and Shoemaker that self-knowledge is required by rationality. On the other hand, Sellars's account of self-knowledge assigns a crucial epistemological role to a causal connection between thoughts and meta-thoughts. It is this causal connection which, after we

acquired it through appropriate conditioning, makes the above principle (3) true; and we saw that in order to have justification for our self-knowledge, we need to know this principle.[14]

The reason why Sellars can assume a necessary connection between rationality and self-knowledge, and at the same time, pace Shoemaker and Burge, assign a crucial epistemological role to a causal connection between thoughts and meta-thoughts is the following: in the case of a subject that has acquired the capacity for non-inferential self-knowledge, this causal connection is, on Sellars's account, embedded in a network of rational connections which confer epistemic authority on meta-thoughts. As Sellars puts it:

> Now the important difference between a person who has *merely* been conditioned to respond to his thought that-p by saying "I have the thought that-p" and a person whose statement "I have the thought that-p" *expresses direct self-knowledge* is *not* that in the latter case the statement *isn't occurring as a conditioned response*. It is. The difference is that in the latter case the conditioning is itself caught up in a conceptual framework. (to Castañeda, 3.4.1961, §13)

So on Sellars's account, our entitlement for self-knowledge is *both* based on a rational connection between first- and second-order thought (the connection expressed in the background inference), *and* on a causal connection between both thoughts (since this causal connection underlies a premise of the background inference). This undermines the opposition between "perceptual" and "non-perceptual" accounts of self-knowledge.

[14] It is true that principle (3) does not itself contain the term "causality". But "conditioned" is a causal notion; and Shoemaker and Burge would clearly count a view on which self-knowledge rests on a link between thoughts and meta-thoughts that can be established by mere conditioning as a version of the perceptual model of self-knowledge, which they aim to refute. (For example, Shoemaker 1994, 288 holds that it is "of the essence" of the first-order state to be accompanied by a higher-order state, and such an essential connection cannot be the result of conditioning.)

But what about Shoemaker's and Burge's arguments? With regard to *Shoemaker's* self-blindness argument, Sellars could simply argue that in the case in which the causal connection in question collapses and self-blindness obtains, rationality collapses, too. (Cf. to Castañeda, 8.12.1961, Ad II.3: "To lose the tendency to have appropriate meta-thoughts is to cease to be rational or lose one's mind".)[15]

By contrast, Sellars's account can be defended against *Burge's* argument quite easily by introducing a further, independently plausible assumption: it is an important feature of our ability for self-correcting reasoning that *if* our meta-thoughts make it seem the case that there is a rational shortcoming at the level of first-order reasoning, we directly take this as a reason for trying to revise the first-order reasoning accordingly. In other words, in the first-personal case, we normally do not make use of the possibility of a brute error as an excusing condition. This policy might simply be a further part of what we learn in acquiring our capacities for self-critical reasoning. It is true that such a policy will lead to cases in which we erroneously attribute to ourselves failures of rationality where what has really gone wrong is the formation of meta-thoughts. But rationality requires anyway, on Sellars's account, that the mechanism of self-ascription works reliably. Therefore, such error cases may be rare enough to not impair the normal exercise of our rational abilities.

3. Self-knowledge and Consciousness

The biggest difference that sets apart Sellars's account from most of the recent accounts of knowledge about occurrent mental states such as thoughts has to do with the relation between self-knowledge and *consciousness*. Most recent accounts agree that our ordinary, non-inferential self-knowledge about such occurrent mental states is restricted to states that are

[15] But see Haag 2001, 324–28, for another view on Sellars vs. Shoemaker. According to Haag, Shoemaker's argument is undermined by Sellars's critique of the Myth of the Given, as it assumes that in perceptual knowledge (unlike in self-knowledge), the object of knowledge is independent of the subject's knowledge of it (Haag 2001, 327). I doubt this point, since the sense in which first-order mental states depend, according to Shoemaker's constitutivism, on the subject's knowledge of them is far stronger than the sense in which external objects of our knowledge depend, according to Sellars's epistemology, on our knowledge of them.

conscious—where the relevant form of consciousness is typically thought to be "phenomenal consciousness" (Block 1995), i.e. qualitative experience. (Some of these authors hold that conscious thoughts have a distinctively cognitive species of phenomenal character, which cannot be reduced to sensory phenomenal characters; other authors hold that thoughts are conscious in virtue of sensory phenomenal character,[16] e.g. of related inner speech or some other imagery: cf. the contributions in Bayne and Montague 2011). Moreover, these authors assign the fact that those states are conscious an important role in the epistemology of self-knowledge. For example, some authors hold that the mere fact that we are in a conscious mental state gives us defeasible justification for believing that we are in that state (e.g. Smithies 2012). Other authors argue that there are background factors which make it the case that whenever we are in a conscious mental state, we have defeasible justification for believing that we are in that state. Such factors may consist, for example, in an a priori knowable principle according to which self-ascriptions of certain mental states that rationally respond to conscious experience are true (Peacocke 2003, 160); or in the fact that in consciousness-based self-ascriptions of present thoughts, we exercise a general rational capacity (McHugh 2012); or, on one version of a "Transparency Theory" about self-knowledge, in an epistemic rule that allows us to move from our experience of inner speech about a certain topic to a self-ascription of a thought about that topic (Byrne 2011).

Unlike such "consciousness-based" views, as we may call them, Sellars does not assign any role to phenomenal consciousness of our thoughts in his account of self-knowledge. Instead, he simply denies that thoughts themselves can have phenomenal character (PSIM 33).[17] And while he grants that thoughts can be reflected in verbal imagery, which *can* have

[16] Some would insist that in this case, only the imagery, not the thought itself, is conscious. But since nothing hinges on this for our present discussion, I shall use the notion of thoughts being conscious "in virtue of" related imagery in a sufficiently broad way to include the possibility that, strictly speaking, only the imagery is conscious—provided it contains enough information about the thought to support (phenomenally) immediate self-knowledge about it.

[17] Sellars's discussion in this context is framed in terms of "qualitative" or "intrinsic character" (PSIM 33f.). What Sellars denies is that thoughts have a qualitative character that can be introspected (PSIM 33). I understand such introspectively accessible qualitative character to be identical with what is nowadays

phenomenal character, he denies that such imagery (or, presumably, imagery of any other form) is necessary for self-knowledge about our thoughts (EPM §47: 178). We can now see that the distinctive structure of Sellars's account of self-knowledge about thoughts is a direct consequence of this view. For while consciousness-based views typically see the route to self-knowledge as an "*ascent*" from a conscious first-order state to a higher-order state that amounts to self-knowledge, this option is not available to Sellars. Instead, his account adopts, as we saw, a reverse order of explanation: he understands self-knowledge in terms of a "*descent*" from the self-ascription of a higher-order state (in premise (1a) of the trans-level inference) to the self-ascription of a first-order state (the conclusion (2) of that inference).

In the remainder of this section, I will consider two possible arguments for a consciousness-based view and against the Sellarsian account, and examine how a Sellarsian can react to them.

1. The first argument is the following. When someone is challenged to defend his beliefs about his present occurrent mental states, it seems natural to reply (although not necessarily in these terms) that it is one's present conscious experience which shows that one has such-and-such mental states. By contrast, a Sellarsian trans-level inference, starting from a premise like (1a), would seem a highly uncommon way to respond to such a challenge. So a consciousness-based account is closer to our actual justificatory practices—or so the argument goes.

I think that this is a natural worry to have with Sellars's account, but not a decisive one. For Sellars could reply that conscious experience can serve as (part of) the justificatory basis for self-knowledge only if some variant of the Myth of the Given is presupposed. More precisely, Sellars could argue as follows. *Either* conscious experience is itself a propositional, belief-like state. Then it can in principle serve as a reason for a self-ascription of a mental

---

normally called the "phenomenal character" of a state, i.e. the "what it is like" of consciously experiencing that state.—When Sellars claims that thoughts lack, while "sensations" possess, phenomenal character (PSIM 32f.), he follows a view that dominated analytic philosophy of mind until recently, and that is itself a legacy of behaviorism: cf. Siewert 2011, 238–42.

state, but it then stands also itself in need of justification. This is not something that any of the above-mentioned views seems to assume. *Or* conscious experience is not itself a propositional state. But since Sellars holds that only propositional states can provide reasons for something else (cf. deVries and Triplett 2000, 76–7), it cannot be in this case the conscious experience *itself* which contributes to the justification of the higher-order belief. Rather, one would need to postulate an additional state that *detects* relevant features of one's conscious experience, e.g. a report like

(5) I am presently having a conscious experience as of a thought that p,

or

(6) My present conscious experience makes it seem to me that I am thinking that p.

This report would have to serve as reason for the self-ascription of the thought that p, and it would have itself to be capable of being justified. But in this case, it is not so clear anymore that the justification for the belief "I am thinking that p" proceeds by ascent from the conscious first-order thought itself. Rather, this justification now seems to start from a propositional state that itself articulates a potential reason for the second-order belief "I am thinking that p", and to this extent speaks *about* that second-order content. It is therefore possible for Sellars to argue that a report of a form like (5) or (6) is itself located at the same level of discourse (the third-order level) as Sellars's own first premise in his trans-level inference,

(1a) The thought that I am thinking that p has occurred to me.[18]

If this line of reasoning is correct, consciousness-based views either need to conceive of the justification for self-knowledge as "descending" from a third-order content to a second-order thought or belief, as does Sellars, or they need to embrace a version of the Myth of the Given on which beliefs can be justified by something else than propositional states. Regardless of whether one thinks that Sellars's attack on the latter option is successful or not, his position

[18] Recall that Sellars uses sometimes "I am under the impression that I am thinking that p"—a formulation that is quite similar to (5) and (6)—as equivalent for (1a).

on self-knowledge turns out to be less extravagant than the comparison with consciousness-based views can make it appear at first sight.

2. The second argument goes as follows. It is a decisive advantage of consciousness-based views over Sellars's theory, according to this argument, that they capture an intuitively plausible connection between self-knowledge and consciousness, to which Sellars fails to do justice. For consciousness-based views imply that ordinary non-inferential self-knowledge of thoughts is restricted to phenomenally *conscious* thoughts. By contrast, unconscious thoughts—e.g. steps in unconscious cognitive processes—are not available for such non-inferential self-knowledge. So consciousness-based views agree (although they make this not always explicit) on the following principle:

(7) A subject S knows non-inferentially at t that she is thinking that p -> S's thought that p is phenomenally conscious at t (either in virtue of some irreducibly cognitive phenomenal character, or of some related sensory phenomenal character).

I think that this principle is eminently plausible. There is only one clear case of thoughts that do not show up in conscious experience (in any of the ways mentioned in (7)[19]): namely, thoughts that are part of the unconscious cognitive processes postulated by cognitive psychology.[20] Our access to such thoughts is, of course, highly inferential in nature. In order to deny (7), one would need to assume that there are thoughts which are like the unconscious thoughts postulated by cognitive psychology in that they do not show up in the stream of consciousness, but which are at the same time also very different in that we can have non-inferential knowledge about them. It seems very hard to identify any cases of thoughts that fit this description.

If (7) is true, a theory of self-knowledge about one's thoughts should be consistent with (7), and be able to explain why (7) is true. But since Sellars assigns no role to phenomenal

[19] Cf. also the note before the last.

[20] Or more precisely, from the Sellarsian viewpoint: by theories in cognitive psychology for which we have not undergone Jonesean training.

consciousness in his account of self-knowledge, he fails to account for the truth of (7). And given his further claims that thoughts (a) lack phenomenal character, and (b) can be non-inferentially known independently of any accompanying imagery, he is even committed to denying (7). So if (7) is indeed plausible, this speaks against Sellars's theory of self-knowledge.

Of course, a Sellarsian may respond to this argument by simply denying the intuitions in favor of (7). But there is another and, I think, more interesting option for the Sellarsian: it is possible to supplement Sellars's account of self-knowledge such that it can actually take into account (7). For Sellars, as we saw, self-knowledge builds on a mechanism through which first-order thoughts trigger corresponding meta-thoughts. Now this idea will have as a consequence a restriction like (7) if it is combined with the view that the occurrence of meta-thoughts which are caused by corresponding first-order thoughts is what makes those first-order thoughts conscious. This is exactly what so-called Higher Order Thought (HOT) theories of consciousness hold. According to such theories, mental states are phenomenally conscious in virtue of being represented by higher-order thoughts that are caused by the first-order states, and that ascribe those first-order states to the subject.[21] If HOT theory is combined with Sellars's account of self-knowledge, the latter becomes apt to do justice to our above observations about the relation between self-knowledge and consciousness. But unlike consciousness-based theories, the resulting view does not see consciousness as a *precondition* for self-knowledge: rather, consciousness and self-knowledge are both consequences from one and the same factor, namely, the production of higher-order thoughts.[22] To be sure, in order to amend her position in this way, the Sellarsian has to abandon Sellars's claim that

[21] The most elaborate version of HOT theory is David Rosenthal's, cf. Rosenthal 2005. Rosenthal himself suggests that his HOT theory can be combined with the framework of Sellars's philosophy of mind (Rosenthal 2005, 304). But as far as I can see, he nowhere relates HOT theory to the account of self-knowledge that Sellars develops in the correspondence with Castañeda.

[22] In order for the premise (1a) in the trans-level inference to be known to the subject, the relevant second-order thought presumably needs to be itself made conscious by a corresponding third-order thought (cf. Rosenthal 2005, 28 on introspection). If that second-order thought is challenged, the third-order thought can be made conscious by a fourth-order thought, so that the corresponding instance of the trans-level inference becomes available, and so on.

thoughts lack phenomenal character. But Sellars's reason for this claim was that such phenomenal character would resist a physicalist explanation (PSIM 31–33). And since HOT theory is precisely meant to give a physicalist explanation of consciousness, this reason becomes obsolete on the proposed view.[23]

Bibliography

Bayne, Timothy and Montague, Michelle. 2011. *Cognitive Phenomenology*. Oxford: Oxford University Press.

Block, Ned. 1995. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18: 227–47.

Burge, Tyler. 1996. "Our Entitlement to Self-Knowledge." *Proceedings of the Aristotelian Society* 96: 91–116.

Byrne, Alex. 2011. "Knowing that I Am Thinking." In *Self-Knowledge,* edited by Anthony Hatzimoysis, 105–24. Oxford: Oxford University Press.

Carruthers, Peter. 2011. *The Opacity of Mind. An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

deVries, Willem and Triplett, Timm. 2000. *Knowledge, Mind, and the Given. Reading Wilfrid Sellars's 'Empiricism and the Philosophy of Mind'*. Indianapolis: Hackett.

Gertler, Brie. 2010. *Self-Knowledge*. London and New York: Routledge.

Haag, Johannes. 2001. *Der Blick nach innen. Wahrnehmung und Introspektion*, Paderborn: Mentis.

Lewis, David. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50: 249–58.

McHugh, Conor, 2012. "Reasons and Self-Knowledge." In *Self-Knowledge*, edited by Annalisa Coliva, 139–63. Oxford: Oxford University Press.

[23] I am grateful to the participants at the Erlangen conference on Sellars for helpful discussion, and especially to Anke Breunig and Stefan Brandt for their thorough comments on earlier versions of this text.

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

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

O'Shea, James. 2012. "The 'Theory Theory' of Mind and the Aims of Sellars's Original Myth of Jones." *Phenomenology and the Cognitive Sciences* 11: 175–204.

Peacocke, Christopher. 2003. *The Realm of Reason*. Oxford: Clarendon Press.

Peters, Uwe. 2014. "Self-Knowledge and Consciousness of Attitudes." *Journal of Consciousness Studies* 21: 139–55.

Rosenthal, David. 2005. *Consciousness and Mind*. Oxford and New York: Oxford University Press.

Ryle, Gilbert. 1949. *The Concept of Mind*. London and New York: Hutchinson.

Sellars, Wilfrid. (1956) 1991. "Empiricism and the Philosophy of Mind." In *Science, Perception and Reality*, 129–94. Atascadero: Ridgeview.

——. (1962) 1991. "Philosophy and the Scientific Image of Man." In *Science, Perception and Reality*, 1–40. Atascadero: Ridgeview.

——. (1967) 1991. "Phenomenalism." In *Science, Perception and Reality*, 60–105. Atascadero: Ridgeview.

——. (1974) 2007. "Meaning as Functional Classification: A Perspective on the Relation of Syntax to Semantics." In *In the Space of Reasons. Selected Essays of Wilfrid Sellars*, edited by Kevin Sharp and Robert B. Brandom, 81–100. Cambridge (Mass.) and London: Harvard University Press.

Sellars, Wilfrid and Castañeda, Hector-Neri, 2006. *Correspondence on Philosophy of Mind*, edited by Andrew Chrucky, online at http://www.ditext.com/sellars/corr.html, accessed February 25, 2018.

Shoemaker, Sydney. 1994. “Self-Knowledge and ‘Inner-Sense’.” *Philosophy and Phenomenological Research* 54: 249–314.

Siewert, Charles. 2011. “Phenomenal Thought.” In Bayne and Montague 2011, 236–67.

Smithies, Declan. 2012. “A Simple Theory of Introspection.” In *Introspection and Consciousness*, edited by Declan Smithies and Daniel Stoljar, 260–95. Oxford: Oxford University Press.