

Externalism and First-Person Authority. Davidson on Self-Knowledge

Miguel García-Valdecasas*
Faculty of Philosophy
University of Oxford
10 Merton Street
Oxford OX1 4JJ

Abstract. Davidson considers self-knowledge a kind of knowledge which relies on no evidence nor yields privileged access to the subject. He also thinks that claims not grounded in evidence bear no more authority than claims based on evidence, questioning thus the idea that self-knowledge exhibits the highest degree of authority. This is implicitly held in his critique of Putnam's externalism, which I briefly examine. Yet the way in which his externalism assigns concepts to natural kinds leaves important questions unresolved and bring about a number of difficulties, one of which sets his view at odds with Wittgenstein's asymmetry of perspectives, which Davidson has allegedly endorsed. Defeasibility is equally committed to an account of the correction of defective self-knowledge which is highly problematical, since it builds on a *cognitive assumption*, namely, the supposition that while subjects enjoy knowledge of their mental states, this knowledge is a sufficient and justified warrant for the validity of their self-ascriptions.

Propositional attitudes such as 'I believe that *p*', 'I want to *p*', 'I desire to *p*' and others are expressions of someone's state of mind for a third-person. For the first-person, in turn, these are thought to be *transparent*, in the sense that they do not bring about any cognitive achievement nor are they acquired in the way in which any perception is acquired. While Davidson¹ seems initially to endorse the transparency thesis, he insists that first-person

* This article has been written with a postdoctoral scholarship co-funded by the Spanish Ministry of Education and the European Social Fund. This document is released under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 Spain License. For a summary of the terms of this license, including a link to the full license, please visit <http://creativecommons.org/licenses/by-nc-nd/2.5/es/deed.en>.

¹ D. Davidson, 'First Person Authority', *Dialectica* 38 (1984), pp. 101-111 (hereafter FPA). 'Knowing One's Own Mind', repr. in Q. Cassam (ed.) *Self-Knowledge* (Oxford: Oxford University Press, 1994) pp. 43-64 (KOM). 'Mental Events', repr. in *Essays on Actions and Events* (Oxford: Oxford University Press, 2001) pp. 207-227 (ME). 'Structure and Content of Truth', *The Journal of Philosophy* 87 (1990) pp. 279-328 (SCT). 'Thought and Talk', repr. in *Inquiries into Truth and Interpretation* (Oxford: Oxford University Press, 2001) pp. 155-70 (TT). 'Three Varieties of Knowledge', in A. P. Griffiths (ed.), *A. J. Ayer. Memorial Essays* (Cambridge: Cambridge University Press, 1991) pp. 153-66 (TVK). 'What is Present to the Mind' in *Grazer*

authority should be understood as a presumption, the presumption that a subject is not mistaken when she attributes to herself beliefs, intentions, desires and other psychological states (FPA, 101). In this respect, Davidson's view of first-person authority might look rather standard, since it rests on the mostly uncontroversial idea that self-knowledge relies on no evidence (p. 103), as many philosophers nowadays will widely accept.

But he equally maintains that self-knowledge is as fallible as knowledge of the world. We are told that 'claims that are not based on evidence do not in general carry more authority than claims that are based on evidence, nor are they more apt to be correct' (p. 103). If the transparency thesis stipulates the highest degree of authoritativeness for a subject on account of the intrinsic safety of self-ascriptions, Davidson could not endorse such a thesis, since the presumption that a subject cannot be mistaken about her own psychological states does not provide her with a higher degree of authoritativeness than that of the third-person, nor is this authority preferable when justified by some sort of introspection.

In short, we must think that self-knowledge is defeasible. In this article, I examine the links between self-knowledge and externalism, the grounds for the defeasibility thesis and the difficulties that such an idea involves. Among these I put forward some points in common between Davidson's account of self-ascription and the arguments of the private linguist, and argue that such dependence sets Davidson's view in need of a careful re-examination.

1. Davidson refines Putnam's externalism

Davidson's analysis of externalism departs from the idea that first-person authority can only be guaranteed by presumption. For him, self-knowledge shares with knowledge of the world its being fallible and apt to correction when occasionally, we need to withhold the authority normally given to a subject. Self-knowledge also shares with knowledge of the world a conceptual framework which is the effect of our cognitive engagement with the world. The right use of this framework is essential both to knowledge of the world and to self-knowledge; its misuse generates in turn a bogus belief which spreads out at any cognitive level. Davidson thinks that when a subject falsely believes *p*, not only can we say

Philosophische Studien 36 (1989) pp. 3-18 (WPM). 'A Coherent Theory of Truth and Knowledge', in *Subjective, Intersubjective, Objective*, (New York: Clarendon Press, 2001) pp. 137-153 (CTT). References to the works of Wittgenstein will be abbreviated as follows: BB – *The Blue and Brown Books*; PI – *Philosophical Investigations*; RPP I – *Remarks on the Philosophy of Psychology*, vol. I.

that her belief is false; the same should also be said of any self-knowledge enclosed with it. The mistake in question does not rest on the fact that, contrary to what the subject thinks, she does not have the belief she thinks she has, but on the mere falsity of p which, once it is taken for true, it overturns the authoritativeness of any second-order belief based on that one. Hence, Davidson draws the conclusion that for self-knowledge to be authoritative, knowledge of the outside world needs careful scrutiny, something which he did in an extensive article aimed to reconcile self-knowledge and externalism (KOM) and save the notion of authoritativeness.

There is a clear major difference between Davidson's externalism and that of Putnam, in whose epistemology he sees no guarantee of safety for the contents of thought. His central disagreement with Putnam concerns the place in which meanings are to be found. Any externalism will agree that concepts derive their contents from natural kinds, while Putnam had asserted that concepts belong in the environment in which they are acquired. Although Davidson accepts that concepts are subject to environmental inputs (KOM, 52), inasmuch as any propositional attitude involves the use of concepts, the correct application of such concepts to sentences should take into consideration their environmental root, which certainly are prolific and hard to trace back; we are told that they are constituted by 'relations to society and the rest of the environment, relations which may in some respect not be known to the person in those states' (p. 52). Unfortunately, the supposition that concepts are *naturally* acquired leaves a subject in need of strong epistemic warrants to ensure that her concept of the transparent liquid which is called 'water' in our planet does not correspond to the concept of 'twater' in some distant planet because of the natural acquisition of such concept in the wrong environment. Paradoxically, Davidson notes that in such cases avowals about 'water' switch their truth-value depending on the planet in which the sentence is uttered. Putnam asserts that the sentence 'here's a glass of water' has different truth-values depending on where the concept of 'water' was learnt. For someone who learnt it on Earth and speaks on Twin Earth not knowing where she is, to say 'here's a glass of water' is false; but to say the same sentence on Earth turns out to be true (p. 57).

Problems with the notion of truth spread out sceptical doubts and lack of first-person authority. If any competent subject is liable to take responsibility for her utterances, an authoritative subject fences off unsafe beliefs by carefully tracing back the original root of her concepts. Now, if self-knowledge has access to one's mental states and this

authoritativeness is grounded in that of knowledge of the world, self-knowledge requires a similar degree of authoritativeness to that of first-order knowledge. Otherwise, the threat of a misfit between these two types of knowledge—between first-order beliefs and the knowledge of those mental states—is too near. Suppose now that a speaker claims to know that p when p has the content ‘here’s a glass of water’, and she asserts it in a context in which, unaware of the difference in composition between ‘water’ and ‘twater’ the speaker means the ‘water’ of Earth. The problem to which Davidson points is this. If her concept was acquired on Earth and the sentence is uttered on Twin Earth, the mistake on p is carried forward in such a way that not only the proposition ‘here’s a glass of water’ becomes false, but also the content of any self-knowledge involved in any sentence of the form ‘I know that this is H_2O ’, thereby making any utterance of this kind self-deceptive. The source of the mistake is certainly the object rather than the psychological verb bound up with it, but if self-knowledge is hostage to the reliability of first-order beliefs, self-ascriptions may not only be false but also unauthoritative, which, as it may be supposed, poses serious consequences about the true reliability of a whole system of beliefs².

The problem has still deeper roots. As A. Woodfield puts it: ‘because the external relation [of concepts] is not determined subjectively, the subject is not authoritative about that. A third person might well be in a better position than the subject to know which object the subject is thinking about, hence be better placed to know which thought it was’ (KOM, 49). Strictly speaking, the subject has missed her chance to take her self-knowledge right, because the concept of ‘water’ which she learnt hinges altogether on the environment in which it was acquired, and seemingly, there is nothing in the relevant concept which tells a subject when the conceptual content is at odds its correspondent natural kind.

Davidson envisages a two-pronged solution to salvage the concept of authoritativeness. The first prong stresses that, contrary to what Putnam thinks, meanings can be in the head even though what I mean may depend on things outside me. The second prong stresses that meanings do not stand in the head as ‘objects’ of which philosophers have traditionally spoken. I will focus on the second prong to avoid one possible misinterpretation of the first one. Davidson considers that the refutation of Putnam’s contention that thoughts are not in the head may bring the ownership of thoughts back to our mind, but it will be irrelevant to

² Davidson extends the consequences only to propositional attitudes—not to what Putnam calls ‘narrow contents’ e.g. pains—, in which concepts seem to be fully known to the speaker (KOM, 49).

our needs of authoritativeness if the real commitments of externalism are not clearly exhibited. So, to get at the bottom of the matter the notion of objects, which *pace* Davidson philosophers tend to see as *entities* standing in certain relationship with the environment (KOM, 62), needs some revision. So, if the identity of objects determines what the thought is about, ‘it must always be possible to be mistaken about what one is thinking. For unless one knows *everything* about the object, there will always be senses in which one does not know what object it is’ (p. 63). The nagging problem, already mentioned by Woodfield, is endemic to any version of externalism in which the conceptualisation of natural kinds is made dependent on the environment, since *pace* Putnam (p. 49) concepts are framed in a way which propositional attitudes cannot scrutinise. Knowledge of the environment seems so the key to the right interpretation of conceptual contents and of their propositional counterparts. Davidson stresses that if, on top of that, one thinks that ‘to have a thought is to have an object before the mind’ (p. 63) the meaning of propositions is going to lie open to a wide range of possible interpretations of which that of the subject is not a privileged one: there will simply be a presumption, rather than a safe warrant, that a subject is not taking herself wrong when she gives propositional expression to her mental states.

The second prong of the argument unfolds the idea that meanings can indeed be in our mind. It sketches how to surrender objects of thought, identified by Davidson as ‘the source of the trouble’ (p. 63), without having to face the kind of objection ordinarily levelled against the private linguist. The private linguist finds concepts in her mind whose meaning is self-restricted, and hence liable to interpretation even by her possessor. Davidson envisages so the notion of the ‘natural history’ of our concepts (p. 63), by which the internal process of conceptualisation turns environmental inputs into mental episodes capable of bringing about meaning. Assuming that contents of thought cannot self-determine their true constituents, a subject’s answer to any question about what she thinks at a time *t* cannot count as a criterion to find it out, not because the subject *does not know* what she thinks, but simply because the ultimate extension of her concepts is not part of the understanding of a sentence. The ‘natural history’ of concepts, which need to be screened for that, is nowhere at hand despite being thoroughly at work since the concepts were first elucidated; its ‘space’, the space of a subject’s concepts, is connected to the environment in which concepts are framed and received into a community. But now, as the

subject is unable to trace back the original history of each one of her concepts, no matter how indefinite that notion of 'history' is, it is superseded by that of objects of thought³.

2. The 'special authority' of self-knowledge

Let us leave the process of conceptualisation and turn to Davidson's account of self-knowledge. We already know of his departing presumption that every subject is fully authoritative to her own mental states in a way in which others are not. This is the source of what Davidson calls the 'special authority' of the subject (FPA, 102), the authority which rests on the fact that, whatever the objective meaning of a sentence may be, the individual subject is the only one who is in a position to authoritatively tell (p. 110). If self-knowledge concerns the authority of psychological verbs such as 'I believe', 'I desire', 'I hope', etc. a way of exploring the limits of this authority is to choose a particular belief, i.e. the belief that 'the house is on fire' to see which features of this sentence can shed any light on its underlying safety. Approaches to this sentence vary according to the aspects epistemically exhibited and built into the corresponding belief. These aspects vary if we look at:

- (i) whether the house is on fire,
- (ii) whether the speaker believes that the house is on fire,
- (iii) and how the fire caused that belief.

Each of these aspects prompts a different question. The response to the threats of externalism deals with the authoritativeness of (iii), which is the kind of authority involved in knowledge of the world. Davidson stresses that this authoritativeness (iii) requires a satisfactory account of (i). We must supply some warrants that the house is on fire to know how we came to believe this. The authoritativeness of the sentence 'the house is on fire' can be seriously compromised if something of the nature of the problem with Twin Earth is on work here. Despite this, Davidson is right and first-person authority is not at risk on the basis that the subject has no special authority towards this (FPA, 102), the sentence is

³ It seems that Davidson takes it for granted that the mere disappearance of objects of thought is a good refutation of incompatibilism, which is the view that externalism undermines our special authority (S. Bernecker, 'Davidson on First-Person Authority and Externalism', in *Inquiry* 39 (1996) pp. 128-32).

simply true iff the house of which we mean happens to be on fire. As to (iii), the problem posed by Putnam's externalism, Davidson thinks that the question is already settled.

As to (ii) Davidson thinks that a subject is not ordinarily mistaken about her *believing* that *p*, which is the source of a special authority not at play in (i) (FPA, 102). Whereas I believe that 'the house is on fire' on the basis of some evidence, I do not 'normally' attribute a belief to myself because of it (p. 103). In other words, the physical events described in 'the house is on fire' (i) have no tendency to establish the psychological or epistemological devices by which these events are coupled with beliefs, desires and other phenomena. In this case, the source of the authority splits up into the reasons why a subject knows *p*, and the reasons why he believes or desires *p*. In this way, nothing from the state of affairs which makes a proposition true makes a belief to be a belief, just as nothing in a desired state of affairs makes a desire to be a desire. The source of first-person authority derives from the fact that by default, any speaker knows what she means, and in ordinary circumstances does not question it.

This suggestion comes in conflict with an idea of Shoemaker's which draws its inspiration in Wittgenstein (BB, 66-67), according to which knowledge of one's mental states is immune to error⁴. Davidson stresses that 'the speaker can be wrong about what his own words mean. This is one of the reasons why first person authority is not completely authoritarian' (FPA, 110). Such a provision purports to be consistent with the Wittgensteinian asymmetry of perspectives (p. 110). By it, we should assume that a subject cannot express herself systematically wrong apart from cases of insincerity, malapropisms and brain damage⁵. Certainly, first-person authority is not awarded for any reason based on what is said, but simply because the supposition that persistent and gross error in the understanding the meaning of one's own words is quite unreasonable. If the speaker were worse-off than the interpreter to understand her own words, she will self-ascribe mental states in the way in which she ascribes mental states to others, namely, treating her own assertions as coming from someone else. Yet this seems to contradict the fact that we endorse the beliefs which we assert or that to assert that *p* is to endorse it, whereas third-person beliefs are not necessarily understood as endorsed by us. In short, there is no

⁴ S. Shoemaker, 'Self-Reference and Self-Awareness' repr. in Q. Cassam (ed.) *Self-Knowledge* (Oxford: Oxford University Press, 1994) p. 81.

⁵ For Davidson 'it makes no sense (...) to wonder whether the speaker is getting generally things wrong' (FPA, 111).

disease of first-person authority by which the subject becomes a third-person to herself. For this reason, Davidson says that ‘in general, the belief that one has a thought is enough to justify that belief’ (KOM, 43). Despite this strategy may seem insufficient to account for the possibility of error, if a speaker is a rational agent who acts sincerely, possible discrepancies between her words and her behaviour are ultimately corrigible in various ways (FPA, 111).

3. Two objections to the two-pronged argument

I will discuss in the following three sections Davidson’s externalism and Davidson’s view of first-person authority to mental states. With regard to his brand of externalism, I will draw out two points. First, under the presumption that his argument against objects of thought is set to correct a problem of authoritativeness in the knowledge of the outside world, we might wonder why, from the thought that meanings are not in the mind—as Putnam seems to think—it is necessary to think that meanings are not effectively in our concepts in the way in which philosophers believe them to be—as objects—, but only in the ‘natural history’ of concepts in a way that looks too far beyond our control.

Undeniably, the concept of natural history of our concepts is a vigorous one, but it is mainly so because it cuts short any attempt of identifying such concepts with *entities* (KOM, 62). However, while the history describes a mere causal relation between entities and concepts, it says nothing of how concepts happens to come about. Hence, it can be argued against the notion of natural history that it seems devised to avoid a possible side-effect of externalism—the appearance of unmanageable objects of thought— rather than to presents us with a constructive alternative to it. To this claim, Davidson might have replied that subjects give meaning not independently from her intentions (SCT, 310), and so, not only throughout the links which concepts hold with the environment. But this is not posed to give us a solution either, since with the disappearance of the objects of thought, an assertion like ‘here’s a glass of water’ as uttered in a foreign environment cannot be more authoritative than it is in Putnam’s account, because I am not answerable to the way in which my concept pick up external objects; it is the environment what takes responsibility for such endorsements.

Thus, the first point is that the concept of natural history can hardly be a solution to the epistemology of concepts. In addition, there is a problem of authoritativeness concerning our mental states, since even though the natural history of our concepts is presented as the best available way of clarifying the ultimate meaning of our propositions, nothing seems to explain why first-person and third-person avowals enjoy different degrees of authoritativeness or how exactly the asymmetry of perspectives holds (FPA, 110).

Therefore, unless the asymmetry of perspectives can shed any light on how the first-person perspective is fully asymmetrical with the third-person perspective, Davidson's objection to Putnam will still be applicable to the natural history of concepts. If I do not although I could, treat my own mental states as I treat others' mental states (KOM, 45), first-person authority stands in need of interpretation in the way in which we interpret other people's utterances⁶. Against this, Davidson might have insisted in that objects do not look like entities any more, which is possibly true. Yet they are still in need of some interpretation of their meaning if mental states are to play any role here, such as to be the mental states in virtue of which we are justified in saying that S believes that *p*. If the disappearance of objects is brought to bear upon mental states, now the question is how such states can be methodologically isolated to help us determine the meaning of particular propositions. This is what the next objection will lay down.

Secondly, there seems to be a question as to the identification of mental states. Davidson noted it when referring to objects as entities whose content is open to interpretation by third-person observers. By getting rid of the traditional notion of objects, Davidson eases the way for his anomalous monism, insofar as lacking in objects of thought, type–type identity theories, not token–token identity theories, remain questionable. This explains that 'the mere fact that ordinary mental states are individuated in terms of relations to the outside world has no tendency to discredit mental–physical identity theories as such' (KOM, 59). So, token–token identity theories are still feasible theories of mind. But now, if mental states are individuated in terms of their relation to the outside world and we want to be able to correlate tokens of thought with tokens of neural activity, how are we going to do this if the correlation thought–tokens, however anomalous, can only be done by selecting a thought-token which can be identified with some neural token? Surely, there

⁶ P. M. S. Hacker, 'Davidson on First-Person Authority', in *The Philosophical Quarterly* 47 (1997) p. 289. The problem is more serious than it looks, because it does not extend only to first-person authority. H.-J. Glock has argued that 'Davidson distorts the concept of understanding by identifying it with interpretation' ('A Radical Interpretation of Davidson', in *The Philosophical Quarterly* 45 (1995) pp. 208 and 212).

may be room for the claim that the correlation token–token does not necessarily mean that the purported identity is anyhow feasible. But the problem is not the feasibility of such an identification but its conceivability, since presently there are no objects of thought in sight to be identified; if we want to determine what a subject thinks at a time *t*, the history of concepts will yield the way in which concepts were previously used, but this is to no avail to validly identify ongoing thoughts. In this scheme we can say that we know *that* we think, but we do not know exactly *what* we think, and that despite Davidson consider our meanings to be ‘so directly before the mind that it is impossible to misidentify them’ (WPM, 3).

The rejection of objects as entities may seem an advisable move to forestall recalcitrant mentalistic readings of the mind, but once mental objects have been removed, it remains to see how the resulting picture fits in Davidson’s anomalous monism. To illustrate it, let us assume that a subject’s belief about the King of France can be individuated by the relation of its constitutive concepts to some particular tokens. Provided that, as it now happens, subjects do not think with objects before the mind, how is the token–token identity of mental and brain states to be understood when a subject’s thought about the King of France does not constitute any kind of entity? Especially, when the paradigm of entities—physical events—is not the right one in order to characterise thoughts. To put it clearer, if a subject’s thought on the King of France does not result in an object of thought such as ‘my thought on the King of France’, the search for a brain token seems as difficult as the attempt of identifying a headache when no physical pattern of pain is present. In this case, the lack of evidence may not satisfy a doctor who is convinced that the headache has some physical counterpart despite evidence to the contrary. In contrast to it, Davidson is not questioning the existence of physical tokens, but that of mental ones, and that despite this jeopardises his purported identification in a similar way as that in which the absence of a physical pattern of pain frustrates the identification of a headache with its brain state⁷.

⁷ It is therefore puzzling to relinquish objects of thought while simultaneously allowing for some tokens to be physical kinds of objects in the hope that they both will turn out to be identical. Equally surprising is that Davidson sees nothing suspicious in the assumption that thoughts exhibit causal relations to neural events, while seeing as problematic their characterisation of thoughts as thoughts. Ayer puts it this way: ‘If mental events cannot be sufficiently pin-pointed to be candidates for subsumption under strict laws, why should it be thought that they can be sufficiently pin-pointed to be identified as causes and effects?’ (A. J. Ayer, *Philosophy in the Twentieth Century* (New York: Random House, 1984) p. 188).

We can put the question in this way. In the absence of an object of thought such as the ‘King of France’, nothing can be singled out as the mental state which carries or constitutes that thought. On the other hand, if I can only identify my thoughts about the King of France by the bare fact that ‘this is what I think’, the fact that *I am thinking this here and now* stands alone as the most likely criterion for identifying mental contents. However, in the absence of objects of thought it is hard to see how the natural history of concepts is apt to give us a criterion which is, if anything, too feeble. Of course, Davidson might have planned the replacement of objects by the history of concepts to show that despite all there is something specific and retrievable about my thoughts, but I am arguing that this is precisely what cannot be the case when no satisfactory definition of a concept has been provided so far.

Giving up the causal history of concepts and maintaining that the bare *thinking* is the required criterion of identity, the problem acquires a different shape. If interpreting one’s own mind is now necessary to identify one’s own thought, the interpretation of a mental state becomes a different mental act in a similar way as that in which first-order beliefs as known by second-order beliefs are different mental acts⁸. In this way, if the identification of a mental state constantly generates some other mental states, allegedly interpretative, the mere occurrence of a thought seems insufficient to verify that my interpretation of what the thought is about is the right one. And the situation worsens when *thinking* is the only available criterion to say why *this thought* corresponds to *this token*. For *thinking* is a too loose criterion to make a good token without providing a large array of qualifications, let alone the fact that the concept is more akin to a type or a hyper-type than to any other thing.

4. An alleged misfit of orders

So far as to the argument of objects of thought. With regard to Davidson’s analysis of the authoritativeness of (i)-(iii) and its connection to self-knowledge, the central quest is that of clarifying how first-person knowledge comes about. There are nevertheless reasons to

⁸ The gap between these two states brings about a problem of connectives. Burge claimed to know how their link works. He says that a second-order thought is self-referentially locked in the first-order content. See T. Burge, ‘Individualism and Self-Knowledge’ in Q. Cassam (ed.), *Self-Knowledge* (Oxford University Press: Oxford, 1994) p. 75. But this thesis has also some difficulties, as ‘I believe that *p*’ seems an statement about a mental state rather than the statement of a fact.

suspect that in this process something has gone awry. As C. Wright⁹ puts it, what Davidson's 'point shows is that if some element of opacity of content is indeed introduced by externalism, it will have to show in the possibility of a different kind of error—not a misfit between the contents of contemporaneous first- and second-order attitudes'. When canvassing externalism, Davidson deals with the issue as if the source of the problem of authoritativeness relies on a misfit between knowledge of the world and self-knowledge. But it seems that such a misfit is not a problem for which externalism needs to be invoked or canvassed. To appreciate it, let us briefly return to someone's belief that there is a King of France as an instance of a proposition known by experience. This belief is obviously false; its falsehood, *pace* Davidson, affects other beliefs which rely on this one, such as the belief that 'The King of France is bald'. Yet it seems as if, in addition to this, the unreliability of the propositional content makes the psychological verb unreliable, so that if self-knowledge becomes defeasible as a result of this, a subject would not only be wrong in believing the King of France to exist, but also in thinking, when she believes it in good faith, that this proposition is true. Even if after a careful reassessment the subject were to question her belief that 'the King of France is bald', coming instead to think that France might not have a king at all, the new belief would be presently considered true in the same way in which the King of France was previously thought to exist. In this context, the switch from 'there is a King of France' to 'France might not have a king' has nothing to do with the reliability of the psychological verb, but with the evidence on which we support our beliefs about the world and refine them. Accordingly, the person who changes her mind on a particular issue does it at the realisation that the belief that *p* is false, not because she did not believe *p* to be true. This is precisely Wright's point here: Davidson's supposed misfit between knowledge of the world and self-knowledge can only be apparent.

Much will be gained with a comprehensive distinction between the nature of authoritativeness in (i-iii) and (ii) is needed then to understand self-knowledge, but such a distinction is marginalised when when it is argued that self-knowledge is purely derivative from knowledge of the world. Bypassing that difference between both kinds of knowledge, self-knowledge ends up as a second-order device engineered to prevent self-deceptive scenarios. However, it is hard to see how the most cautionary approach to epistemology

⁹ C. Wright, 'The Problem of Self-Knowledge (II)', in *Rails to Infinity: Essays on Themes from Wittgenstein's Philosophical Investigations* (Cambridge, MA: Harvard University Press, 2001) p. 347.

can ever justify some reasoning on these lines: A believes p , p is false and A does not know that is false; but despite the fact that A claims to believe p , A does not believe p .

5. The argument of the defeasible internal authority

This naturally prompts the question of how authoritative a subject can be with respect to her words and the mental states to which these words give expression. Davidson asserts that part of the job of a competent speaker is to set her mind on the understanding of her words. The best she can do to be authoritative is ‘to be *interpretable*, that is, to use a finite supply of distinguishable sounds applied consistently to objects and situations he believes are apparent to his hearer. Obviously the speaker may fail in this project from time to time; in that case we can say if we please that he does not know what his words mean’ (FPA, 111). Davidson is not clear—and I think that this omission is crucial—as to whether subjects interpret their mental states or rather their propositions, and if the latter, how mental states are standardly turned into propositions. Textual evidence (ITI, 277; FPA, 110) points primarily to the idea that what is interpreted are the speaker’s utterances, but if that suppositions includes the fact that unauthoritative utterances can derive from ‘authoritative’ thought as I have suggested, there is no particular reason to endorse any of the options. The problem of unauthoritativeness may lie on the cause, or in the transition from the cause to the effect, and insofar as the question is not settled, the burden of proof would lie on Davidson.

Davidson insists that ordinary speech undergoes problems of authoritativeness. While each speaker is presumably left to contend with her own problems, we may ask what defeasibility entails for first-person authority. If it entails that the speaker may occasionally utter sentences at variance with her mental states, it looks as if language, including meaning and speech, were an articulated series of mental events which could be variously interpreted, or whose interpretation is not always straightforward¹⁰. Davidson seems to see the reduction of the mental to the physical unfeasible on account of the inexistence of strict psychophysical laws capable of predicting and exploring mental events (ME, 208), but on the other hand, a language potentially redescribable in terms of mental events forcibly demands such a reduction. Therefore, urged to embrace this reduction we could wonder

¹⁰ This seems to be Davidson’s view of interpretation in a language. See H.-J. Glock (1995) p. 208.

whether anything other than the non-existence of strict psychophysical laws stands in its way. In a scenario in which we did not have to do physics by way of laws, such a reduction might appear feasible. In that case, provided we would contrive a scientific method for the identification of tokens not based on laws manageable to non-scientists, anyone would be entitled to invoke her mental events as evidence of the kind of thoughts which she is having. Anyone could point to her mental events to show that they are the same thing. However, if despite all the subject were unable to single out the relevant token or to identify the wrong one too often, how could we justify self-knowledge?

The defeasibility of the belief that p is not the question; the question is the defeasibility of believing that I believe that p . As Davidson notes, one's own words occasionally fail to reflect one's thoughts once cases of insincerity, malapropism or brain damage have been ruled out. Any speaker utters words whose meaning is not the one that she naturally intends, and this causes misunderstandings in ordinary speech. From here, Davidson concludes that the expression of thoughts is an intrinsically fallible process, which shows that 'we do not always have indubitable or certain knowledge of our own attitudes' (FPA, 103). A straight consequence of the defeasibility of self-knowledge is the defeasibility of first-person authority. On that assumption any speaker can mistrust her own words even when they are in accord with one's own thoughts even when they are in accord with her beliefs and for all she knows they happen to be correctly expressed, that is, even when there is no reason to suspect the opposite. Once defeasibility affects one's propositions, there is no reason to think that the failure cannot be carried all the way down so that all mental acts become potentially deceptive. If this is true, the theory should take into account that incidentally, a speaker might self-ascribe beliefs that she does not hold; and so, someone's belief that 'the earth has existed for a long time' might not necessarily be the best expression of her beliefs however convinced she may be that that proposition expresses them accurately. If we were to press the speaker to think this again, what is suggested might become true; she might realise that her words are not the best expression of what she really believes. Yet provided that there is not any warrant to self-ascribe the putative belief, the speaker does not have the leverage to dismiss beliefs which she does not endorse either. On closer scrutiny it might also turn out that the belief that 'the earth has existed for a long time' is the accurate expression of what she thinks, contrary to our supposition. And if asked to confirm this, the speaker replies: 'yes, that is what I think', would there be any way to find out about the speaker's real mental state? Or more

significantly, who would be entitled to judge it? After all, what guarantees that the speaker's mental states do not *contain* opposite or contradictory beliefs to those expressed?

If the correspondence between one's beliefs and one's words is as Davidson proposes, that that seems to be the implication. By interpreting her mental states differently, a subject might discover that despite she believes to believe *p*, in fact she does not believe it. To eschew such a consequence, the objection can be levelled that one needs not evidence of what one believes in situations in which it does not seem reasonable to doubt about one's own words, and that is what Davidson seems to suggest by stressing that self-knowledge relies on no evidence (FPA, 103). But even in that case, accuracy in speech would not be the rule but the exception; despite having the certainty that one holds a particular belief, a subject might still be willing and entitled to check up again the correspondence between her words and her beliefs. Yet the systematic comparison between beliefs and words opens up to an infinite regress, because for any imaginable circumstance in which a subject is totally in command of her speech, the possibility of a misfit between beliefs and words cannot be ruled out, and if that is true, the general warrant to think that when I *believe p* I do not rather *hope* or *desire* that *p*, and vice versa, lacks any real basis. As a result, we should drop the idea that self-attribution can be immune to this kind of error.

6. The cognitive assumption and the private linguist

Davidson thinks that the fact that the 'self-attributer does not normally base his claims on evidence or observation' does not explain first-person authority, because sometimes that knowledge is based on evidence and observation (FPA, 103). Subjects do not use their thoughts as evidence of what they think, but nothing prevents us saying that subjects know or have cognitive access to their mental states. This assumption, deeply embedded in Davidson's account of authoritativeness, is called by Hacker¹¹ the *cognitive assumption*, because it assumes that psychological verbs derive first-person knowledge from our mental states. 'There is a presumption—as an unavoidable presumption built into the nature of interpretation—that the speaker usually knows what he means' (FPA, 111). 'Our thoughts are 'inner' and 'subjective' in that we know what they are in a way that no one else can' (TVK, 165). The speaker can think so that 'he has a warrant for thinking that he has said

¹¹ P. M. S. Hacker, p. 287.

something true in saying ‘I *V* that *p*’¹², and that her warrant is her own self-knowledge. Knowledge of such states justifies for someone who believes *p*, that for all she knows, her self-attribution is correct, and unless a malignant spirit induces the subject to error, knowledge of one’s own mental states would be *justified* when, being the case that we *V* that *p*, we sincerely aver to *V* that *p*. Here, I will side-step the question of whether the cognitive assumption rests on an observational or introspective model of knowledge, models which Davidson rejects (FPA, 104), but which the cognitive assumption can hardly hold back so long as nothing specific is said about how self-knowledge operates.

Justification in self-knowledge brings us back to the transparency thesis, which claims that self-knowledge is not any cognitive achievement. Now, it is undisputable that the cognitive assumption implies rather the opposite, namely, that something other than *p* is additionally known when we claim to *believe* that *p*. Considering this, I want to suggest that Davidson’s account of authoritativeness shares some ground with arguments familiar to the private linguist in Wittgenstein’s *Philosophical Investigations*. The private linguist sets himself to privately name sensations which he previously has identified, such as a sensation of pain, and feels justified to assert that he has a pain on the basis of that particular sensation for which he thinks that his feeling is a warrant. By contrast, for Davidson, first-person authority builds on the fact that the speaker ‘usually knows what he means’ (FPA, 111) and ‘cannot wonder whether he means what he says’ (p. 110). We should presume that his self-attribution is truthful; and that her belief that she thinks that *p* provides the required warrant (KOM, 43).

Whereas the comparison between Davidson’s self-interpreter and the private linguists is appropriate in some sense, in others it is not. The private linguist stands alone in that a private sensation of pain, if genuine, cannot be a propositional attitude; it ought not to display any propositional content to remain inherently private. However, Davidson’s self-interpreter thinks that he is justified in calling them ‘this’ (PI §267). In identifying a sensation, Wittgenstein makes the private linguist reckon that he can be right or wrong (PI §270). For Davidson, as shown above, any mental elucidation is equally vulnerable to error. Both the private linguist and Davidson’s speaker cannot be fully authoritative in self-ascribing mental states, nor is their cognitive position in the whole any safer. Wittgenstein

¹² P. M. S. Hacker, p. 292.

stresses precisely this much to the discomfort of the private linguist, who remains convinced that when he says ‘I am in pain’ I am at any rate justified *before myself*’ (PI §289).

Ever since Wittgenstein it has been standardly accepted that mental states and their expression are not connected by the interpretation of an inner event, that is, that the expression of such a knowledge is a rule-governed practice founded on public criteria. Subjects do not express their mental states in ways only recognisable to their possessor. Thus, we do not assume that someone who breaks in tears can be misinterpreting her own mental states or pretending to feel grief. She could not be blamed for having cried while being in a state whose natural expression is to cry, because crying is not the activity we choose to do in order to grieve, but a natural expression of grief. Likewise, Wittgenstein warns about the idea that a cry expresses pain in the same way as a proposition expresses thought (PI §317). Taking the utterance of a proposition as stemming from a *mental sensation* leaves out an important difference between sensations and thought. Wittgenstein asserts that ‘one can mistrust one’s own senses, but not one’s own belief’ (PI II, X, 190e), for which we do not want to suggest that sensations and thought are tailored in the same way. To maintain, however, that we know our mental states and can be wrong about them is different from the claim that self-ascriptions are as defeasible as any other assertions, for while self-ascriptions do not always succeed in expressing someone’s thought, this does not entail that self-knowledge is inherently fallible and their expressions are unjustified.

At the heart of this picture lies the belief that expressions are fully detachable readings of thought to which they are accidentally connected, presumably resembling the independence that the private linguist grants to the identification and expression of sensations. Once this scheme is applied to mental states, thoughts appear as *mental occurrences* whose content is laid out throughout the specific expression of that experience (RPP I §105) and is essentially private (RPP I §109). On such a view, meaning might well be associated with mental states and expression with sentences, whereas language might be the combination of both into a single process. Of course, the independence of both phenomena will allow the speaker to say: ‘I am not merely saying this, I mean something by it’ (PI §507) as an instance of the way in which thinking can be reduced to an accompanying act of speech, ‘a process which may accompany something else, or can go on by itself’ (PI § 330).

Wittgenstein's reply to the private linguist stresses that 'when we speak of someone's having given a name to pain, what is presupposed is the existence of the grammar of the word "pain"; it shows the post where the new word is stationed' (PI §257). The private linguist requires some command of the grammar of 'pain' to find out and name his sensations. In the same way, Davidson's self-interpreter presupposes the grammar of psychological verbs such as to 'know' when it is said that 'I know that *p*' is a fallible self-expression and that the subject needs to be given a chance of setting her self-knowledge right. In turn, the suggestion that I can deceive myself in thinking that 'I know *p*', not because *p* but because I am not in a mental state of knowledge, discloses a particular kind of mistake. If any answer to the question about my mental contents presupposes the grammar of 'to know', I should not say that 'I know I was wrong' or 'I know I was right' on pain of circularity. Therefore, Davidson's self-interpreter seems neither authoritative nor competent to say whether she believes *p* or not, because the use of 'to know' is subject to the general defeasibility of any belief. This is why the private linguist refuses to think whether he is sufficiently authoritative to identify his own feelings. Wittgenstein ironically adds that for him 'this question does not matter the least' (PI §270).

Interestingly, Davidson might have agreed with Wittgenstein that the private linguist requires cognitive verbs to give expression to his feelings. He thinks that 'to notice', 'wonder', 'know' and 'remember that' are psychological verbs which can be redescribed in terms of 'belief' (TT, 156-7). The description of such verbs in terms of belief is a way of providing public criteria to some phenomena which otherwise might appear to be enclosed with privacy. Yet Davidson seems to be less aware of what the acceptance of this idea entails for his cognitive assumption, and this is what I have tried to spell out.

The solution to the puzzles of the private linguist is not to grant special authority to self-knowledge. Wittgenstein implicitly endorsed the transparency thesis in arguing that the realisation of our mental states involves no knowledge. We cannot make any philosophy out of the idea that a self-expression is a statement about the subject. If the person who says to hope *p* does not describe a hopeful state of affairs, what is she describing? Is the hope to *p* a description of a state of mind independent from the state of affairs described by *p*? Davidson affirms precisely this. He thinks that 'when I say 'Jones believes that snow is white' I describe Jones's state of mind directly: it is indeed the state of mind someone is

in' (TT, 167). Although he attributes the relevant mental state to a third-person, we have seen that there is no reason not to apply this argument to the first-person, for which it also holds. In this way, 'I believe that it is raining' is some sort of self-description, the description of an inner event which carries a proposition. If this person were questioned about her belief, the answer might follow these lines: "At bottom, (...) I am describing my own state of mind—but this description is indirectly an assertion of the fact believed".

—As, in certain circumstances I describe a photograph in order to describe the thing it is a photograph of. But then I must also be able to say that the photograph is a good one. So here too: "I believe that it's raining, and my belief is reliable, so I have confidence in it".

—In that case my belief would be a kind sense-impression' (PI II, X, 190e¹³), when 'sense-impression' can be read here as advancing some knowledge of the world.

The words 'I believe it's raining' do not describe the mental state of the speaker, and therefore, they are not an expression of self-acquaintance or of self-knowledge. They are simply the description of a state of affairs as known by a subject and not an expression of any *mental feeling*. This seems what Davidson leaves out when considering that the authoritativeness of the first-person entails the cognitive assumption and that this knowledge is not immune to error. Therefore, Davidson's theory of self-knowledge is mentalist, as it conceives of mental states as interpretable processes in need of an account of the mental similar to that of the private linguist to feature the *special authority* of the subject. It cannot be ruled out that the starting point, the connection between externalism and first-person authority which comes into play with Davidson's criticism of Putnam, is the reason why the mental appears now as 'entities before the mind' (KOM, 62).

8. A doorway to scepticism

If what I suggest is correct, Davidson's account of first-person authority has common grounds with the private linguist's arguments, perhaps much of what Davidson would be prepared to recognise. His account is left in a difficult position inasmuch as thought can be seen as a private and intrinsically defeasible activity to which only a private interpretation can be tentatively advanced. The subject is then left in a situation in which neither her utterances nor her beliefs are adequate criteria to determine the content of her own

¹³ P. M. S. Hacker, pp. 292-3.

thoughts. If that is the case, Davidson's arguments cannot dispel the internal threat of sceptical doubts. The presumption that one knows that one thinks, or the *cognitive assumption*, needs to spell out how such knowledge is possible and can be presumed of a subject other than by supposition, since only subjects who trace back the provenance of their concepts and carefully check the correspondence between words and beliefs qualify for authoritativeness. The cognitive assumption might not compromise his idea that 'beliefs are by nature generally true' (CTT, 153), because they are indeed, but it does not provide us with grounds to keep the sceptical doubt well apart from the psychological.

[7,821 words]