Published in German in: *Deutsche Zeitschrift für Philosophie* 62(5), 2014, pp. 937-962.

The Development of the Authority of the First Person

Johannes L. Brandl
University of Salzburg

Abstract: This paper defends an expressivist approach to explaining the epistemic authority with which self-ascriptions of mental states are made. Following other recent defenses of cognitive expressivism, it is assumed that avowals (e. g. 'I am feeling pain now') have both a descriptive and an expressive function. Since these functions can come apart, expressivists face the problem of false avowals, which threaten to undermine any claim to epistemic authority. In response to this problem it is suggested that expressivists may draw a distinction between sincere avowals and avowals in the strict sense, and then argue that observing this distinction is part of a linguistic competence which is a precondition of reflective self-consciousness. From these plausible assumptions one can see that no appeal to introspection as a form of privileged access to the mind is needed for explaining the authority of mental self-ascriptions.

Keywords: first-person authority, cognitive expressivism, avowals, reflective self-consciousness

This essay explores the question of what role the development of reflective self-consciousness plays for the authority of the First Person. By reflective self-consciousness I mean a higher-level awareness in the form of cognitive acts that relate to one's own sensations, thoughts and feelings. Persons who have developed such meta-awareness are thus in possession of an epistemic quality called the authority of the First Person. But why is that? What gives this epistemic relevance to reflective self-consciousness?

Studies on the development of self-consciousness tend to avoid answering this epistemological question by focusing on the ability of self-reference. Self-reference together with other representational abilities is fundamental for the development of a self-concept.[1] Thus, one focusses only on questions pertaining to the acquisition of a self-concept and separates them from questions about the authority of the First Person. That this is a popular move becomes manifest in

[1] See e.g. Bermudez 1998.

the debate on immunity to error through misidentification, initiated by Sydney Shoemaker[2] The epistemic aspects of this debate only concern the acquisition of a self-concept. The question of the authority of the First Person is set aside.

However, it is not at all clear that one can decouple the question of the authority of the First Person from the question of the development of self-consciousness. Traditionally, the subject of self-consciousness has always been guided by an epistemological interest. Both rationalists and empiricists assume that our self-concept is more than a sum of 'I'-thoughts, the justification of which can be discussed separately. Self-consciousness is traditionally a 'consciousness of consciousness' with the special feature that it is the source of immediate self-knowledge. On these premises there is no room for questioning the epistemic relevance of self-consciousness, and there is also no room for doubting the authority of the First Person. Entertaining such doubts would mean doubting the very existence of a reflective self-consciousness, and that would be downright absurd.

But the authority of the First Person is *not* beyond all doubt. It therefore needs some other explanation, so I will argue, why there is indeed a connection between the development of self-consciousness and this epistemological problematic. If it is not our self-knowledge, as the traditional view holds, what else could explain the authority of the First Person to one's own mind? Since Wittgenstein we have known a possible alternative: the authority of the First Person could also be a *linguistic* phenomenon. It could derive from the fact that we learn to express our feelings and thoughts in the form of *avowals*.[3] In my opinion, this approach has three critical advantages: First, following this line of explanation means that we no longer need to appeal to a problematic model of privileged access to one's own mind. Secondly, it can be argued that expressive abilities are not only fundamental for the authority of the First Person but also for the development of reflective self-consciousness. And third, we see in light of the limited credibility of expressive utterances why there can be legitimate doubts about the authority of the First Person.

[2] See Shoemaker 1968.

[3] See Wittgenstein 1953.

Even so, there is still more than one way to go in developing a notion of authority of the First Person that does not rely on the concept of privileged access.[4] Among the possible options available, Wittgenstein's approach still stands out as the most promising, as one can see from a number of studies that have recently taken up and systematically developed this approach.[5] While my reflections are largely based on these works, there are some difficulties left that require a somewhat different approach.

Section 1 elaborates further the basic problematic and formulates the central questions that a theory of the authority of the First Person has to answer. Section 2 presents the basic idea of a concept of first-person authority based on what I call "Descriptive Expressivism". To see why the idea needs further qualification, I introduce in Section 3 the problem of false avowals. Section 4 engages with Rosenthal's view that this problem has the potential to undermine any expressivist analysis of self-knowledge. In response to Rosenthal, and as a solution to the problem, I introduce in Section 5 a distinction between sincere and genuine avowals. In Section 6, where I return to the questions raised at the outset, I finally defend the view that the linguistic competence from which the authority of the First Person originates can be regarded also as a decisive factor in the development of our reflective self-consciousness.

1. Questions about the authority of the First Person

There can be no doubt that the development of higher cognitive abilities is promoted, if not made possible in the first place, by learning a language. But it remains a matter of controversy which of the higher cognitive abilities only develop with language and which are already acquired prior to learning a language. What about the reflective self-consciousness and the authority of the First Person in this regard?

From an epistemological point of view, there is no reason to associate these phenomena from the start with language acquisition. The starting point here is rather the thesis that a subject

[4] Worth mentioning here is Rorty's attempt to replace the Cartesian concept of infallibility with a conventionalist concept of incorrigibility (Rorty 1970), but also Andre Gallois' attempt to explain the authority of the First Person by appealing to claims about the transparency of the mind (Gallois 1996). Both approaches neglect the important role that avowals play in this context.

[5] See Falvey 2000, Bar On and Long 2001, Finkelstein 2003, Bar On 2004, 2010 and Green 2007

is directly conscious of the contents of her own mind—that is, possesses a consciousness of its own consciousness. For this consciousness to exist, a subject need not be able to articulate her meta-consciousness linguistically. But, then, the question arises as to how direct self-knowledge not bound to linguistic abilities could develop. Cartesians seek a metaphysical explanation for this in the form of an ontologically primitive relationship in which each subject stands to its own *cogitationes*. Empiricists, on the other hand, might explain the direct acquaintance with our own minds in terms of a process similar to perception, namely introspection. Although both of these two classical models of self-knowledge are highly controversial, they have lost none of their appeal, as the ongoing debate on them shows.[6] In a modernized form, these models also form the basis for cognitive explanations of our self-knowledge through representation processes that register, like an internal monitor, other mental representations and feed this information into processes at a higher level. The constant assumption here is that an internal monitoring device is also supposed to work independently of linguistically acquired skills.[7]

However, one should take this continuity as an indication of an emerging consensus that there is direct self-knowledge which is not tied to language. There is not only Wittgenstein's already mentioned expressivist theory that contradicts this, but also the very different view of Donald Davidson, whose essay "First Person Authority" illustrates the profound disagreements on this matter in philosophy.

Davidson addresses the question of the authority of the First Person from the beginning in the context of philosophy of language. That is how he justifies this move:

> "[The question of the authority of the First Person can be asked] either in the modality of language or epistemology. For if one can speak with special authority, the status of one's knowledge must somehow accord; while if one's knowledge shows some systematic difference, claims to know must reflect the difference. I assume therefore that if first person authority *in speech* can be explained, we will have done much, if not all, of what needs to be done to characterize and account for the epistemological facts." (Davidson 1984, 102, my emphasis)

[6] See Gertler 2011, Hatzimoysis 2011, Smithies and Stoljar 2012

[7] See Newen 2005; Vosgerau 2009

The aforementioned advocates of direct self-knowledge would of course contradict Davidson's claim. From their point of view, there exists an asymmetry between self-attributions and attributions to others precisely because we know of our own mental states in a direct way. What could be wrong with that? For Davidson, what is at stake here is a methodical question about the direction of explanation. If there is direct self-knowledge, then it makes sense to say that the utterances with which we express that knowledge inherit their peculiar epistemic quality. But there is also another way how one can explain the "systematic difference" within our knowledge, as Davidson puts it. One can analyze the authority of the First Person also first "in speech" and then develop on that basis a theory of self-knowledge.[8]

But what recommends this other way preferred by Davidson? It is difficult to find a substantive argument for this in Davidson's reasoning.[9] All one finds are remarks to the effect that any appeal to a consciousness of one's own consciousness at this point becomes "empty", i.e. without explanatory power (Davidson 1993, 249); or the claim that we are dealing here with a philosophical myth which Davidson calls the "myth of the subjective" (Davidson 1988). What exactly constitutes this myth, however, is hard to say. There is mention of the fact that the myth is rooted in metaphysical dualism, Cartesian claims to infallibility and the transparency of the mind. Beyond that, however, the myth boils down to the claim—made without justification—that the authority of the First Person is based on a particular form of self-knowledge. Thus, nothing has been gained for the question of the factual correctness of this claim.

Since Davidson's argumentation tends to remain empty-handed at this point, I will place my critique of the idea of immediate self-knowledge on an expressivist thesis. The following crucial difference here is noteworthy: Davidson considers it appropriate to discuss the question of the authority of the First Person first for statements in which a speaker assures that he has certain beliefs, desires, hopes, or intentions. "What holds for the propositional attitudes," says Davidson, "ought, it seems, to be relevant to sensations and the rest." (Davidson 1984, 102). The expressivist point of view, by contrast, recommends to discuss the question of authority first for avowals with which speakers express their sensations and feelings. What we can learn about

[8] According to Davidson, however, there wouldn't be much left to do here if everything had already been done with the first step to describe the epistemological facts, as he thinks.

[9] See Davidson's reply to Thöle 1993.

reflective self-consciousness from the expression of sensations and feelings can then guide us in dealing with other mental states we are reflectively aware of.

To facilitate the further discussion, I would now like to provide a list of the central questions and motivate the order in which I will deal with them. To state my questions, I distinguish the following two theses, the relationship of which needs to be clarified:

(Thesis A) A speaker possesses direct self-knowledge of the feelings and sensations that he ascribes to himself due to a privileged access to these experiences.

(Thesis B) A speaker who ascribes some sensation or feeling to himself does so with the authority of the First Person.

Thesis (A) corresponds to the classical models of self-knowledge through direct acquaintance or through introspection. Compared to this, thesis (B) is a much weaker thesis, which only claims that there is an authority of the First Person, but says nothing about where it comes from. Thesis (A) therefore includes thesis (B), but not vice versa.

The first question I ask will then be this:

(Question 1) Does it make any sense at all to accept thesis (B) without also accepting thesis (A)?

This question concerns first of all the *concept of* authority. Can we *understand* this term without implicitly presuming what is made explicit in thesis (A)? If the answer to this question is positive, as I will argue, we get to a second question:

(Question 2) Why should thesis (B) be true if thesis (A) does not hold true?

Unlike Davidson, I think that for answering this question more needs to be done than stating principles for the interpretation of linguistic utterances from which it follows that ascriptions to oneself are largely true. For a principle of charity applies to whatever assertions speakers make quite generally. That is why I regard an expressivist analysis of self-attribution as a more

promising alternative to models relying on immediate self-knowledge. My third question is then the following:

(Question 3) Why is it that avowals that are made with the authority of the First Person form a category of utterances distinct from assertions in general?

As we shall see, the distinction between theses (A) and (B) still plays a role here. Because such utterances would just be ordinary assertions that express our self-knowledge if the authority of the First Person with which we use avowals were founded in the immediate self-knowledge of the speaker. If, on the other hand, avowals are *not* ordinary assertions, then this speaks against the assumption of an underlying self-knowledge, which is expressed in such expressions. I hope that the reasoning here makes apparent how closely the first three questions are linked.

Yet there is another question that I will deal with in this essay. My fourth question concerns the role that the development of ex self-consciousness plays in this context:

(Question 4) Why do speakers need to have reflective self-consciousness so that they can utter avowals with the authority of the First Person?

As long as one sticks to thesis (A), the answer to this question is simple. As soon as one drops thesis (A) and only holds on to thesis (B), however, answering that question becomes more difficult. Whether it is worth taking these difficulties upon oneself remains to be seen.

2. Authority without privileged access

In his book *The Opacity of the Mind* (2011), Peter Carruthers describes human consciousness as a kind of veil that conceals the cognitive processes that determine our behavior. According to Carruthers, our mental self-knowledge is largely based on mechanisms of self-interpretation. The many examples of confabulation and self-deception show that these mechanisms are not particularly reliable. We cannot lift the veil of consciousness any more than we can cast a direct and unmediated glance into the consciousness of others. In both cases, we work with more or less

well-confirmed theories about possible reasons for action that the veil of consciousness hides from us. According to Carruthers, the idea that there is a clear asymmetry between self-attribution and attributions of mental states to others is therefore based on an error that must be explained empirically.

At first sight, this is a position that is at least as diametrically opposed to the Cartesian view of the human mind as Davidson's view. But that is not really the case. For there are two points in which Carruthers account of the opacity of mind coincides with the Cartesian thesis of transparency. First, Carruthers assumes that there is a "sensory self-knowledge" free of interpretation (Carruthers 2011, 72). And secondly, he maintains that in those cases where there is an authority of the First Person, this authority is founded in a cognitive mechanism that is tantamount to a privileged access to one's own mind. While such a privilege is the rule for Descartes, it is the exception for Carruthers. This—admittedly small—concession to those who subscribe to the idea of immediate self-knowledge raises an interesting question: Does one really have to make this concession if one wants to do justice to the intuition that, at least in the case of sensory experiences, there is an authority of the First Person?

Unlike Carruthers, I think that no appeal to the idea of privileged access needs to be made even in the case of sensory states. I admit that it is far from clear how opaque our consciousness really is. Perhaps many of our self-attributions are indeed dependent on self-interpretations that are influenced by prejudices and therefore unreliable. But that may not be the only reason why a naive concept of authority fails the empirical test. In any case, we need—and I agree with Carruthers on this point—a concept of first-person authority that is consistent with all available empirical data.

If the authority of the First Person is understood as a linguistic phenomenon, one arrives at a concept of first-person authority that has a good chance of satisfying this requirement. I will first present this concept in the form of a definition and then discuss it in detail:

(DEF 1) A speaker S has the authority of the First Person with respect to an utterance *x*, if and only if:

(i) utterance *x* describes a mental event or state of which S is aware at the time of utterance.[10]

(ii) S has the linguistic competence with respect to *x* that one would expect from a normal speaker.

(iii) the fact that S expresses *x* is a *prima facie* reason to believe that *x* is true, provided that conditions (i) and (ii) are satisfied.

The basic idea of this definition is very simple: recognizing the authority of the First Person means that we are generally disposed to believe a speaker what he says without invoking further evidence. If we judge a self-attribution to be made with competence, this is evidence enough. This does not mean that self-attributions are free of error, nor that they possess the highest possible degree of certainty. So self-attribution can also be wrong.[11] However, these are always errors which can be attributed to special circumstances, i.e. exceptional cases. *Prima facie*, the fact that an avowal has been expressed is a good reason to assume that it is true.

Although all this seems intuitively very plausible, it is worth considering the conditions of DEF1 individually. To illustrate this, let us consider the common case of a patient who comes to the doctor with knee pain. Following the usual routine, the physician asks the patient to demonstrate in which position his knee hurts. The patient starts a knee bend, stops, and then says pointing to his left knee:

(1) Now it hurts here.

It does not matter if the patient can say anything more about his pain. Already his statement (1) can be regarded as a minimal description of a sensation, so that condition (i) is fulfilled in any case.[12]

---

[10] Strictly speaking, in this definition it would not be necessary to limit the authority of the First Person to mental phenomena *at the time of utterance.* Those who report experiences that they can vividly remember can also claim the authority of the First Person for these reports. For the sake of simplicity, however, I am sticking to the usual limitation to what is currently conscious.

[11] Such an exceptional case—the case of pain anxiety—I will treat later.

[12] It would be a misunderstanding to think that condition (i) could only be fulfilled if there is another cognitive state in addition to the feeling of pain, namely the belief 'I am in pain now'. I will show in Section 3 that this is not necessary.

However, we can only be sure of this if condition (ii) is also fulfilled. Let us therefore imagine a case in which the verbal competence of a speaker is doubtful. Suppose there existed a highly reliable lie detector that could determine whether someone was sincere or deliberately telling the truth. Let us also assume that someone whose sincerity is beyond doubt claims to be in pain, but otherwise behaves normally, cheerfully and calmly, as if he had no pain at all. Then we would be as confused as in the case of a child who has not yet understood what the word 'pain' means. Likewise, in the case of an adult who apparently says 'I have pain' out of pleasure or mood, without expecting help or pity, there is the suspicion that there is a misunderstanding or linguistic abuse. Condition (ii) is meant to take care of the fact that such cases cannot be used as counter-examples against the authority of the First Person.

Something similar can be said about condition (iii). In this case, too, the complication introduced by this condition is needed for taking care of certain borderline cases. For example, a swindler who only pretends to be in pain can take advantage of the fact that his message is usually considered credible. This means that we must also grant the authority of the First Person to the successful swindler, because he uses it to arouse compassion or deceive someone. But what about a person who is generally known to be prone to imagining or pretending to imagine pain? We cannot deny a hypochondriac or a notorious swindler that he sometimes tells the truth. One could even hold the view that there are painful sensations even without a clear physiological cause, which is why even an imaginary pain could be regarded as something painful. What is in doubt again in this case, however, is the linguistic competence of such persons. They can still say they are in pain, but can they really mean what they are saying? That, of course, depends on what is meant by 'mine' here. Surely their utterances cannot do all that a normal speaker does when he expresses pain. Hypochondriacs and notorious swindlers are not taken seriously. Condition (iii) shows why this is an exceptional case and not a counterexample to the thesis of the authority of the First Person.

But what about the many confabulations and self-delusions that Carruthers refers to? Do these examples show that most of our self-attributions are based on self-interpretations and therefore have no special first-person authority? If this were so, we would have to add to the definition as a further condition that the speaker's utterance "must not be based on any self-interpretation".

The normal case of the patient describing his pain to the doctor shows why such a condition would be problematic. Because how should a doctor know whether a self-interpretation is involved or not? If this condition were added, it would not even be possible in a very ordinary case to decide whether someone has the authority of the First Person or not. The practice is that the doctor believes what a patient says, even if the possibility of self-interpretation is not excluded. For a skeptic like Carruthers, this may be a naive attitude. For me it merely shows that, there is no conflict—*at a conceptual level*—between saying that someone engages in a self-interpretation and saying that he or she has the authority of the First Person.

Further difficulties arise, however, if we move on to the question of *why* speakers have this authority. Once the idea of privileged access comes into play, there is undoubtedly a conflict between ascribing to someone authority based on privileged access and saying that his or her self-attributions are based on self-interpretations. But what can we conclude from that? For Carruthers it follows that in those cases—namely in the case of sensory self-knowledge—privileged access in the form of some metacognitive mechanism must be assumed because the authority of the First Person cannot be denied in such cases. But there is also another conclusion that can be drawn at this point. The conflict at hand might also show that the authority of the First Person does not have to be based on granting a person privileged access to her mind. It may just as well derive from the *expressive* function of her self-ascriptions. To this we can add: If self-attributions do not have this expressive function, then they arise from self-interpretations and therefore cannot claim any first-person authority. My task now is to make plain why this is the correct conclusion.

3. Avowals

Considered as a linguistic phenomenon, the authority of the First Person is tied to self-attribution as a particular form of verbal expression. The term ‘self-attribution’ is too general for characterizing the present case precisely. Utterances that have a special expressive function, deserve a special name. That is why the term “avowal” has been coined as term for utterances with a special expressive function. My usage of the term follows this tradition (see Gasking 1966).

The choice of this term goes hand in hand with the obligation to explain what the special function of avowing consists in. The task is not an easy one, as we will see in a moment. Questions arise about the semantic properties and rules of use for such expressions, including whether avowals can be both true and false, and what a sincere avowal comes to. The basic question we must begin with, however, is why avowals are not ordinary assertions.

Let us consider the passage in Wittgenstein's *Philosophical Investigations* which is regarded as the *locus classicus* for the expressivist analysis of avowals (PU §244). Wittgenstein asks himself the question: "How do words *refer* to sensations?" For him, the difficulty of the question lies in the fact that it is not clear how "a person learns the meaning of the names of sensations". Wittgenstein then goes on to describes the following possibility:

> "Words are connected with the primitive, the natural, expressions of the sensation and used in their place. A child has hurt himself and he cries, and then adults talk to him and teach him exclamations and, later, sentences. They teach the child a new pain-behavior.
>
> "So you are saying that the word 'pain' really means crying?"—On the contrary: the verbal expression of pain replaces crying and does not describe it." (PU §244)

To what extent can linguistic expressions be compared with natural expressive behavior, such as crying, screaming or groaning, but also laughing or cheering? Wittgenstein suggests that words like "pain" are learned by children by adding these words to their pain behavior. This can easily be misunderstood as a behaviorist thesis. However, that is a delicate issue that I want to postpone to the end when I will look at some related claims of Wittgenstein. Even so, we can exclude at this point the possibility that the word "pain" does not refer to pain itself, but only to a certain pain behavior. This clarification, which Wittgenstein explicitly makes in the above quote, still leaves open a crucial question: What does it mean to say that a linguistic term "replaces" the cry and does not "describe" it?

It is a popular view among interpreters of Wittgenstein to ascribe to him the following thesis:

(SEX) Avowals have a purely expressive function and no descriptive function at all, in the same way in which natural expressions of our feelings like moaning, crying or laughing, lack such a function.

The acronym "SEX" is short for "simple expressivism". This is the term that Dorit Bar On uses for the view that Wittgenstein is famous for (Bar On 2004, 228). Whether Wittgenstein actually was a simple expressivist, however, is controversial (see Finkelstein 2003).[13] In any case, there are two good reasons that speak against the thesis taken as such.

One reason is that uttering an avowal is a speech act, and how could a speech act substitute for behavior such as crying or moaning? It is at least premature to claim that linguistic utterances can express a feeling in the same way as a behavioral response does. Bar On therefore proposes that one should not compare the expressive function of avowals with a natural (causal) form of expression. Rather, one should analyze it as a form of expression of its own. In reference to Sellars she calls it the "action sense of expressing" (Bar On 2004, 216).

The second reason speaking against (SEX) goes deeper. Avowals behave logically and syntactically like ordinary sentences. One can infer from the avowal 'I have pain' the conclusion 'Someone has pain'; one can syntactically link the avowal with other sentences, e.g. by using them as an antecedent of a conditional; and one can contract avowals with other sentences to make a complex assertion. For instance, by contracting 'I have pain' and 'He has pain' one may assert: 'We both have pain' or 'He has pain and so do I'. All this suggests that avowals, like other assertions, have a descriptive function. They do not only express but also *describe* sensations, feelings or thoughts. A scream doesn't do that. In this respect, avowals are markedly different from anything that has a natural expressive function (see Geach 1965, Wright 1998, Bar On 2004).

Yet, so far there is no reason to back down from the thesis that avowals are no normal assertions due to their specific expressive function. This is a position that one may be inclined to defend, as we will see in the next section. But the objections to *simple expressivism* do not recommend that position yet; they only invite us to examine more closely what the specific expressive function of avowals might be.

Basically, there are two ways to determine an expressive function more precisely. One can refer to the *genesis* of their function or one can characterize more precisely *what* is expressed. If one speaks of a "natural expressive function", one follows the first path by referring to a causal origin of their function. The aforementioned "action sense of expressing", which Bar On prefers,

[13] Note that Wittgenstein in the quoted passage only says that the word 'pain' does not describe *crying*; he does not say that it does not describe pain. (SEX) finds therefore no supported in the above quotation.

also gives a genetic answer to our question. By contrast, my suggestion is to take the other route which aims at determining the specificity of avowals in *what* they express, not *how* they do it.

Let us therefore compare an avowal with an ordinary assertion from this point of view. What is the difference, if we interpret the statement 'I have pain' either as an avowal (A) or as an ordinary assertion (B)? In the first case, we can say:

(A) S is the verbal expression of a state of consciousness M in which the speaker is at the time of the utterance.

If, on the other hand, we interpret the same utterance as a common assertion, we should say:

(B) S is the verbal expression of the speaker's *belief* that he is in state M at the time of the utterance.

Admittedly, the difference between (A) and (B) is difficult to draw when we consider the example of expressing pain, because in this case it is not so easy to separate the belief of having pain from the sensation of pain. Let me therefore change the example and use for M the mental state of secretly falling in love. We can assume that such a feeling can remain hidden also from the subject, if someone is unsure whether he (or she) is actually in love or not. It could therefore be that someone has to be convinced by his friends that he is actually in love because he is not ready to admit it to himself. As long as this state persists, his utterance 'I am (not) in love' could not be an avowal. It could at best be interpreted as an assertion that expresses a belief to which others have encouraged him. Strictly speaking, someone in this situation should therefore say 'I *think* I'm in love'.

That way of explaining the difference between avowals and ordinary assertions has some obvious advantages. We are now no longer obligated to answer the delicate question of whether, and in what sense, the expressive function of avowals counts as natural. In addition to that—and this is an even greater advantage—there is no longer any reason to deny that avowals have a descriptive content. The special feature of an avowal, we can now say, is that it *both* describes *and* expresses a mental state of the speaker. In this respect, I concur with Kevin Falvey that the view called "cognitive expressivism" is to be preferred over "non-cognitive expressivism":

> "[We must] avoid the errors of non-cognitivist expressivism, and take seriously the idea that avowals have an expressive function, while retaining the natural and compelling idea that the sentences used in making them are capable of truth or falsity, and presented as true, so that such avowals also have the pragmatic status of assertions" (Falvey 2000, 73).

A crucial point that needs further explanation here concerns the "pragmatic status" of an assertion. It could simply mean, as Falvey suggests, that assertions are statements capable of being true or false. But it could also convey a stronger thesis, according to which an avowal retains *all* the functions of an ordinary assertion while taking on the specific function of avowing. On this view, an avowal of the form 'I am in M' would fulfil at least *three* functions. It would (1) describe the state M, (2) express the state M, and additionally (3) express a belief with the content [I am in M].

In principle, nothing counts against attributing all three functions to a single statement. Yet, one should not turn it into a condition that holds for *every* avowal.[14] We should not forget about the elementary avowals of children for whom such a complex analysis would certainly be inadequate. When a child says 'I am thirsty', it may describe and express the state M, but not a belief with the content [I am in M]. We need a formulation of the expressivist view of avowal that leaves room for such elementary avowals. I call it the thesis of *Descriptive Expressivism*:

(DEX) An avowal is the simultaneous description and expression of a mental state which may or may not also express that the speaker believes that he is in the state that he attributes to himself at the time of the utterance.

The question of how avowals differ from ordinary assertions is thus answered, but only preliminary. As I will argue below, thesis (DEX) needs to be restricted to what I will call "genuine" avowals. In the next section I discuss the problem that leads to this restriction.

[14] Here I agree with Bar On when she behaves cautiously towards a dual expression thesis (see Bar On 2004, 307ff.).

4. The problem of false avowals

Avowals express precisely those states which a speaker ascribes to himself by means of this avowal. As we have seen, one can advocate an expressivist thesis without determining whether it has a natural (causal) or another kind of expressive function. A question that an expressivist cannot avoid, however, is the following: Are there also false avowals? If his answer is yes, the expressivist needs to tell us how it can be wrong what the speaker says, if an avowal expresses exactly the mental state that the speaker attributes to himself.

The problem might not seem a pressing one at first. It is clearly possible that someone signals through his non-verbal behavior that he is in pain without this being the case. All the more it should be possible for someone to say 'I am in pain' even though it is not true. Neither in the one case nor in the other is it impossible for someone to behave as if he were in a state M without being in M.

As David Rosenthal has made clear, however, a problem emerges if one ascribes both a descriptive and an expressive function to avowals (see Rosenthal 2010). Consider the following example of a false avowal. It is a well-known phenomenon that patients sometimes claim to feel pain out of for fear of impending pain. The so-called *dental fear* is a case of such "pain anxiety". Patients show all signs of pain, although they should be painless from a medical point of view. For example, a tooth is treated whose nerve is dead, or a local anaesthetic has been administered to prevent any sensation in the treated area. Even if an error can never be completely ruled out in such cases, the probability that the patient only imagines the pain is usually much higher. It is therefore natural to think that the patient feels something, e.g. the vibration of the drill, which is in this case a painless sensation that prompts a pain reaction (see Rosenthal 2010, 109ff.).

A straightforward principle that one can follow in analyzing such examples would be this: What does not exist cannot be expressed either. If the pain is only imagined, the patient can only act as if he had pain, but he cannot *express* his (non-existent) pain. Rosenthal does not stop here, however, when he draws the following conclusion from this example:

> "The possibility of detectable factual error in such cases suggests that we should see remarks such as 'I am in pain' and 'I think it's raining' as straightforward reports of individual's mental states." (Rosenthal, 2010, 109).

At first, this sounds as if Rosenthal wants to eliminate any difference between the perspective of the First Person and the perspective of the Third Person. His suggestion seems to be that avowals—whether true or false—are only reports that someone makes about himself from the Third Person's perspective. However, this is not what Rosenthal actually means. To avoid the problematic connotations of the term "reporting", we might put his claim better in this way:

(R) Avowals express the first personal belief of a speaker that he is in the mental state described therein at the time of the utterance.

That makes it clear what Rosenthal denies. He denies that there is any significant difference between an avowal and an ordinary assertion. Because expressivists like Bar On search for such a difference, they think that in an act of avowing no "brute error" is possible (see Bar On 2004, 200). However, as the example of pain anxiety shows for Rosenthal, a simple error is quite possible when patients are confused between two different sensations.

Rosenthal thinks that he also has an explanation of what leads expressivists astray. It is the fact that avowals can exhibit a special form of equivalence relation with ordinary assertions. They can be "performance conditionally equivalent" with assertions, as Rosenthal says (Rosenthal 2005, 274ff. and 2010, 118). By denying only one term of such an equivalence relation, a so-called Moore paradox can be generated. This happens, for instance, if asserted at one and the same moment:

(1) It's raining.

(2) I don't think it's raining.

A similar inconsistency occurs when you say 'Ouch, hurt!' in the same breath and then add 'I have no pain'. In this case, too, one gets caught in a pragmatic contradiction, because the avowal 'I have no pain' expresses a belief that one cannot have when one is in pain, just as one cannot have the belief not to think it is raining when one believes that it is raining. It thus seems that Rosenthal's analysis can be generalized, and that the performative contradictions that arise here are essentially connected with the expressive function of avowals.

What, then, is the problem that one who subscribes to the expressive view of avowals has to face here? Rosenthal's argument makes plain that avowals behave pragmatically in a special

way, and that no further consideration regarding their expressive function is necessary for explaining how a pragmatic contradiction arises (see Rosenthal 2010, 118ff.). One may take that as a methodological objection to the expressivist thesis. In addition, however, there is another fundamental assumption to which expressivists are committed at risk here. According to Rosenthal's analysis, it is no longer possible to explain the authority of the First Person as a linguistic phenomenon independent of the idea of privileged access to one's own mind. The reason for that is a connection that needs explanation between the pragmatic phenomena involved and the authority of the First Person. Hence, the following question arises: Why do these contradictions occur only when one of the two statements is made with the authority of the First Person? Answering this question leads us back to the idea of privileged access, as the following remark by Rosenthal shows: "When one describes oneself as believing something or as being disposed to do so, it's on the basis of one's subjectively unmediated access to the states." (Rosenthal 2010, 119).

However, it is precisely by trying to explain the authority of the First Person as a linguistic phenomenon that one wishes to avoid such recourse to the idea of privileged access. If we continue to adhere to this project, we must therefore solve the problem of false avowals differently from what Rosenthal proposes.

5. Sincere and genuine avowals

Without going any further into the analysis of pragmatic contradictions, which is a topic of its own, I simply acknowledge that Rosenthal is certainly right on one point: the fact that two statements have the same performance conditions does not mean that they have the same expressive function. But it does not exclude this either, and this is enough to defend the expressivist thesis against Rosenthal's argument.

In the case of an ordinary assertion, we know when such an assertion is considered sincere. It is sincere when the person making the assertion believes what she claims. By contrast, someone who believes the opposite of what she claims is dishonest. If avowals were ordinary assertions, their sincerity could then be explained in exactly the same way and the case would be

settled. However, since avowals are not ordinary assertions according to the thesis of expressivism, an alternative explanatory approach is necessary.

What we need here is a distinction between *sincere* and *genuine* avowals. *Sincere* avowals can be false, but not *genuine* avowals. That one has to make this distinction can be illustrated by the performative equivalence of 'Ouch!' and 'I am in pain' already mentioned. This equivalence should be maintained even if the speaker suffers from pain anxiety and misrepresents his own mental state. What we cannot deny the patient is his sincerity when he says he is in pain. Accordingly, we cannot describe his moaning as insincere. But how can this be if he has no pain at all, as we presume?

Instead of explaining the sincerity of an avowal by the *presence* of a corresponding belief, an expressivist can use the *absence* of the opposite belief as a reason. We thus arrive at the following indirect definition of sincerity:

DEF2 The avowal by which a speaker S attributes a mental state M to himself is sincere if S does *not* believe at the time of the expression that he is *not* in pain.

The double negation in this definition may at first be irritating. What could this mean other than that S must believe to be in M? I will give three reasons why this indirect definition is more than just a logical gimmick.

The first reason for defining the sincerity of avowal in this indirect way is an empirical one. As I have already mentioned, we must not forget the elementary avowals that children express at a stage of their cognitive development in which they have sensations, but no reflective self-consciousness yet. That this phase exists is by no means incompatible with a higher-order theory of consciousness such as the one advocated by Rosenthal. Even if, according to this theory, sensations are only conscious when a self-attribution in the form of a higher-level thought occurs, such thoughts do not necessarily imply any reflective or introspective self-consciousness (see Rosenthal 2005, 47ff.). Granting Rosenthal this point, we still need to consider whether children who do not yet have reflective self-consciousness can be sincere or dishonest in their avowals. That is now the crucial question on which it depends whether one can grant them an

authority of the First Person or not. Defining the concept of sincerity as suggested above in an indirect way, allows us to raise this question meaningfully.[15]

The phenomenon of pain anxiety, as previously described, can also serve as an argument here. I have already mentioned that a subject suffering from pain anxiety can hardly be denied to be sincere when she claims to feel pain. However, what remains unclear is whether one should say that the patient *wrongly believes* to feel no pain. How can one decide whether such a belief exists or not? The indirect definition of sincerity saves us from this difficulty. The safer option here is to say that a patient in this situation clearly does *not* believe that he is *not in* pain. So, his avowal is sincere, unlike that of a swindler who knows (or at least believes) that he has no pain, but still claims it.[16]

That we can now solve the problem of false avowals differently from what Rosenthal proposed counts in my view as a third and decisive reason in favor of an indirect understanding of sincerity. For Rosenthal, we have seen, false avowals are simply assertions that express a false belief of the speaker. That makes false avowals to be genuine avowals, just as false assertions in general are genuine assertions. My intuition says otherwise here. I would not call the avowal of someone who tries, but fails to express pain that he does not have at all, a genuine avowal, even if the utterance is sincere. I therefore submit that we should define the concept of a genuine avowal in this way:

DEF3 A speaker's avowal is a *genuine* avowal only if it is both *sincere* and *true.*

On this definition, the utterance of a false avowal is a "deficient" speech act and should therefore not count as a genuine avowal. These utterances do not fulfil the expressive function they claim

[15] Also, one might bring into play here again the difference between the perspective of the first person and the perspective of the third person. For a report from the Third Person perspective to be sincere, the person giving that report must believe that her report is true. By contrast, if someone expresses from the perspective of the First Person that she feels pain, for example, she does not have to believe in the truth of her assertion in order to be sincere. It suffices if she does not believe the opposite of what she says.

[16] Interesting in this context is the as if play by children. The fact that children are not dishonest could be explained as follows: In the context of an "as-if" game, someone can say that they are in pain while convinced that they are not in pain, because the context of the game ensures that there is no deception of others. If, on the other hand, sincerity is understood in the weaker sense of DEF2, the following explanation offers itself as an alternative: Kids don't think they're in pain when they say they're in pain in a "as-if" game. But since they also do not believe that they do not have pain, they can pretend that they are sincere.

to fulfil. They do not express the states the speaker attributes to himself. Only genuine avowals perform this double function. This was the reservation I mentioned at the end of the previous section: We must restrict the truth of (DEX) to genuine avowals.

One should not think though that a similar restriction needs to be made now in defining the concept of the authority of the First Person. That is not necessary, because to say that an avowal is made with the authority of the First Person only means, as we have seen in Section 1, that its utterance by a competent speaker is a *prima facie* reason to think that things are as the speaker says. That still remains true even if it turns out that it was a false avowal.

6. Authority and reflective self-consciousness

This brings me back to the opening question of this essay: What role does reflective self-consciousness play in the development of the authority of the First Person? I am now considering this question on the presumption that we can take the authority of the First Person to be a linguistic phenomenon as specified in (DEF1). Furthermore, I assume that the thesis of Descriptive Expressivism (DEX) provides us with the explanation of where this authority comes from. Against this background, let us now address the fourth question in my list:

(Question 4) Why do speakers need to have reflective self-consciousness so that they can utter avowals with the authority of the First Person?

On the view defended here, the answer to this question derives from the specific expressive competence associated with the use avowals. That competence is something like the common root for both: the reflective self-consciousness and the authority of the First Person. If that is so, it must be possible to show that the kind of behavior that we deem to be sufficient to attribute reflective self-consciousness to a person, is also sufficient to attribute the authority of the First Person to that person, and vice versa. Only when we understand both as essentially related—the reflective self-consciousness and the authority of the First Person—does become clear what special role our expressive competence comes to play here: it fulfils exactly the same function traditional epistemology assigns to what it takes to be our direct self-knowledge.

Let us compare this result with the reflections of Davidson and Wittgenstein considered earlier. The critical question that Davidson asked was this one: Should one discuss the authority of the First Person "with regard to language" or "with regard to epistemology"? The difficulty here turned out to be how Davidson can justify his preference for the first alternative over traditional models of immediate acquaintance and introspection. Now we can see where this difficulty comes from. It comes from the fact that Davidson has no explanation for the fact that the authority of the First Person cannot exist without reflective self-consciousness. Such an explanation can be found, however, if we add the expressive function of avowals.

So, let us look again at Wittgenstein's remark about how we learn sensory expressions. His view seems to be that children learn expressions like 'pain' simply by adding the utterance of such words to their behavioral repertoire. They say 'pain' or 'something hurts' instead of just crying out. Nevertheless, Wittgenstein did not want to advocate a behaviorist thesis in the sense that the word 'pain' only refers to the crying, not the sensation. So far so good. But one can still find a behaviorist thesis in this remark. Perhaps Wittgenstein thought it was a simple conditioning process that underlies this learning process: A child learns to say 'pain' so that it can be comforted; to say 'thirst' so that it can drink; to say 'hunger' so that it can eat; etc.[17]

The exegetical question, whether Wittgenstein meant it that way or not, is not of my concern here. What interests me is this: Why is the language learning process not such a conditioning process? Arguably, this is because children learn mental vocabulary *not only* by learning to use avowals. This learning process is part of a complex social interaction that takes place reciprocally: At the same time, when children learn to express their feelings they also learn to understand the expressions of others. By grasping the intentions of someone who says 'Ouch!' or 'Thirst', they also understand what triggers such expressions. At the age of two they learn to empathize with others (see Bischof-Köhler 2011).

[17] Eine Variante eines solchen Konditionierungsprozesses bilden 'Aufstiegsroutinen', wie sie R. Gordon mit folgendem Beispiel veranschaulicht. "There is a [...] way of training children to use the form of a self-ascription of belief. It would be possible to take a child – say, a two-year old – who can make just a few simple assertions, like 'It's raining', and in one easy step train her to sound *very* sophisticated. Just get her to preface her assertion, optionally, with 'I believe'. She'll then say, for example, 'I believe it is raining'. (Gordon (1995), 61). Solche Routinen generieren gemäß Gordon zunächst zu 'uncomprehending ascriptions': "They will not have learned that they believe it is raining, or even that they have beliefs. As far as they are concerned, they are just parroting a formula before saying what they *really* mean to say, namely, that it's raining." (ibid.).

We consider a child to be linguistically competent only when he or she has sufficiently mastered the dual role of speaker *and* interpreter of other persons. Without this competence, there is no authority of the First Person, as definition (DEF1) makes clear. Even in the case of a parrot, Whatever amount of conditional learning a parrot may have experienced, we would not allow it to express its feelings with the authority of the First Person. The parrot may say 'I'm thirsty' at the right moment, but he still would not understand what he is saying, at least not as we would expect from a normal speaker. It is the social interaction with other language users that uncovers the descriptive content of these utterances.

Here is a final argument to drive home this conclusion: Would it be conceivable for someone to learn to understand the meaning of mental words, if he was never reflectively aware of his own sensations, feelings or thoughts? I think we can rule this out as impossible. But we can also exclude, as I have argued, that someone understands such expressions when they are uttered by others without using them himself with the authority of the First Person. This would contradict the interactive way in which mental expressions are learned. It is inconceivable that a speaker, in the course of language acquisition, only understands that others have certain sensations, thoughts and feelings, without also understanding that he himself has or can have similar sensations, thoughts and feelings. The linguistic competence thus turns out to be the link between two cognitive achievements. On the one hand, that competence allows speakers to express their mental states with the authority of the First Person; on the other hand, it is the critical condition that allows subjects to acquire a reflective self-consciousness.

At no point in this account do we have to mention an immediate self-knowledge. Let me therefore, at the end, modify a famous analogy of Wittgenstein to highlight the critical aspect of this result. The beetle in the box, which only I can see, is not part of the language game, Wittgenstein says. The thing in the box has no significance and is therefore "irrelevant." (PU §293). What could the "thing" be that we can cut short when we apply Wittgenstein's analogy to the expression of sensations? Certainly not the sensation, nor the knowledge, that one feels something. What we can cut out is *direct self-knowledge*, i.e. the traditional notions of introspection and private access to one's own mind.[18]

[18] I would like to thank Frank Esken for his valuable comments on earlier versions of this essay.

literature

Bar-On, Dorit (2004), *Speaking My Own Mind. Expression and Self-Knowledge* (Oxford: Clarendon Press).

——— (2010), 'Avowals: Expression, Security, and Knowledge: Reply to Matthew Boyle, David Rosenthal, and Maura Tumulty', *Acta Analytica,* 25, 47-63.

Bar-On, Dorit and Long, Peter (2001), 'Avowals and first-person privilege', *Philosophy and Phenomenological Research,* 62, 311-35.

Bermúdez, José Luis (1998), *The Paradox of Self-Consciousness* (Cambridge, MA: The MIT Press).

Bischof-Köhler, Doris (1991), 'The development of empathy in infants.', in M.E. Lamb and H. Keller (eds.), *Infant Development: Perspectives from German-speaking countries* (Hillsdale: Lawrence Erlbaum), 245-73.

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

Davdison, Donald (1984), 'First Person Authority', *Dialectica,* 38, 101-11.

----- (1988), 'The Myth of the Subjective', in M. Benedikt and R. Burger (eds.), *Consciousness, Language and Art* (Vienna: Edition S. Verlag), (reprint in: D. Davidson: *Subjective, Intersubjective, Objective*, Oxford: Clarendon Press 2001, 39-52).

----- (1993), 'Reply to Bernhard Thöle', in R. Stoecker (ed.), *Reflecting Davidson.* (Berlin: De Gruyter), 248-50.

Falvey, Kevin (2000), 'The Basis of First Person Authority', *Philosophical Topics,* 28, 69-99.

Finkelstein, David H. (2003), *Expression and the Inner* (Cambridge, Mass.: Harvard University Press).

Gallois, André (1996), *The World Withoug, the Mind Within* (Cambridge: Cambridge University Press).

Gasking, Douglas (1966), 'Avowals', in R. J. Butler (ed.), *Analytic Philosophy* (Oxford: Basil Blackwell), 154-69.

Geach, Peter (1965), 'Assertion', *Philosophical Review,* 84, 449-65.

Gertler, Brie (2011), *Self-knowledge* (London: Routledge).

Gordon, Robert M. (1995), 'Simulation without introspection or inference from me to you', in Martin Davies and Tony Stone (eds.), *Mental Simulation. Evaluations and Applications* (Oxford: Blackwell), 53-67.

Green, Mitchell S. (2007), *Self-Expression* (Oxford: Clarendon Press).

Hatzimoysis, Anthony (ed.), (2011), *Self-Knowledge* (Oxford: Oxford University Press).

Newen, Albert (2005), 'Why do we enjoy an authority of the First Person?', in A. Newen and G. Vosgerau (eds.), *Knowing your own mind. Self-knowledge, privileged access and the authority of the First Person* (Paderborn: Mentis), 165-83.

Rorty, Richard (1970), 'Incorrigibility as the mark of the mental', *Journal of Philosophy,* 12, 399-424.

Rosenthal, David M. (2005), *Consciousness and Mind* (Oxford: Clarendon Press).

——— (2010), 'The mind and its expression', *Self, Language, and World* (Atascadero: Ridgeview), 107-26.

Shoemaker, Sydney (1968), 'Self-reference and self-awareness', *Journal of Philosophy,* 65, 555-67.

Smithies, Declan and Stoljar, Daniel (eds.) (2012), *Introspection and Consciousness* (Oxford: Oxford University Press).

Thöle, Bernhard (1993), 'The Explanation of First Person Authority', in R. Stoecker (ed.), *Reflecting Davidson.* (Berlin: De Gruyter), 213-47.

Vosgerau, Gottfried (2009), *Mental Representation and Self-Consciousness* (Paderborn: Mentis).

Wittgenstein, Ludwig (1953), *Philosophical Investigations.* (Oxford: Basil Blackwell 1958 (= PU).).

Wright, Crispin (1998), 'Self-Knowledge: The Wittgensteinian Legacy', in Crispin Wright, Barry C. Smith, and Cynthia MacDonald (eds.), *Knowing Our Own Minds* (Oxford: Clarendon Press), 13-45.