POLISH JOURNAL OF PHILOSOPHY

VOLUME VI/2/2012

ANDRZEJ BIŁAT
MARTIN F. FRICKE
ALEXANDER J. GILLETT
JEAN-PIERRE MARQUIS
CLAYTON PETERSON
DANIEL RÖNNEDAL
ALFREDO TOMASETTA

Jagiellonian University

POLISH JOURNAL OF PHILOSOPHY

The *Polish Journal of Philosophy* (*PJP*) is a peer reviewed journal publishing valuable contributions on any aspect of philosophy.
PJP invites the submission of articles, book reviews, discussions, responses, and notices from professional philosophers. It is particularly interested in publishing contributions by new authors who pursue their academic development.
The principal aim of *PJP* is to promote the best of the living Polish philosophical tradition, especially the Lvov-Warsaw School of analytic philosophy and the phenomenological school of Roman Ingarden.
PJP is edited and published at the Jagiellonian University, Kraków. For more details and abstract information for the *Journal*, see also our website www.pjp.edu.pl.

CONTACT

POLISH JOURNAL OF PHILOSOPHY
Institute of Philosophy
Jagiellonian University, Kraków
Grodzka 52
PL-31-044 Kraków
tel./fax 0048 124224916
e-mail: editor@pjp.edu.pl

SUBSCRIPTION

Polish Journal of Philosophy is published twice a year, in May and November. The annual subscription rates worldwide for 2013 and 2014 are as follows:

Libraries and institutions: $100 (75EUR/55GBP)
Individuals: $30 (22EUR/16GBP)
Polish Libraries and Institutions: $30 (22EUR/16GBP)

All prices include costs of shipping. For more details please contact the Editorial Office of *PJP*. Inquiries about advertisement rates, exchange, and other information, should be addressed to the Editorial Office of *PJP*. E-mail: subscription@pjp.edu.pl

Cover designed by Ewa Bolińska

Printed in Kraków.

ISSN 1897-1652

ISSN 1897-1652

POLISH JOURNAL OF PHILOSOPHY

Volume VI, No. 2 (2012)

Jagiellonian University

Kraków

This conclusion is in accord with the common, but often not properly justified, view of modern philosophers that epistemological silence with regard to the external world is the most general kind of skepticism possible.

References

Ajdukiewicz, K. (2004). *Zagadnienia i kierunki filozofii. Teoria poznania. Metafizyka.* [*Issues and Directions in Philosophy. Theory of Knowledge. Metaphysics*]. Kęty-Warszawa: Wydawnictwo Antyk-Fundacja Aletheia. (First edition 1949).

Russell, B. (2007). *The Problems of Philosophy*, New York: Cosimo. (First edition 1912).

Woleński, J. (1995). Logika sceptycyzmu [*The Logic of Skepticism*]. In: J. Paśniczek et al. (Eds.), *Między logiką a etyką. Studia z logiki, ontologii, epistemologii, metodologii, semiotyki i etyki. Prace ofiarowane Profesorowi Leonowi Kojowi*, (pp. 179-184). Lublin: UMCS.

POLISH JOURNAL OF PHILOSOPHY
Vol. VI, No. 2 (Fall 2012), 15-32.

Rules of Language and First Person Authority

Martin F. Fricke
Universidad Nacional Autónoma de México

Abstract. This paper examines theories of first person authority proposed by Dorit Bar-On (2004), Crispin Wright (1989a) and Sydney Shoemaker (1988). What all three accounts have in common is that they attempt to explain first person authority by reference to the way our language works. Bar-On claims that in our language self-ascriptions of mental states are regarded as expressive of those states; Wright says that in our language such self-ascriptions are treated as true by default; and Shoemaker suggests that they might arise from our capacity to avoid Moore-paradoxical utterances. I argue that Bar-On's expressivism and Wright's constitutive theory suffer from a similar problem: They fail to explain how it is possible for us to instantiate the language structures that supposedly bring about first person authority. Shoemaker's account does not suffer from this problem. But it is unclear whether the capacity to avoid Moore-paradoxical utterances really yields self-knowledge. Also, it might be that self-knowledge explains why we have this capacity rather than vice versa.

Can the "rules of language" explain first person authority? In this paper, I discuss three ways of relating language and first person authority: Bar-On's neo-expressivism, Wright's constitutivism and Shoemaker's ideas about Moore's paradox and self-knowledge. All three accounts can be seen as attempts to explain first person authority as a consequence of rules that govern our language. But is it a coherent strategy to explain such authority by reference to rules of language? I shall argue that it is not. Bar-On's and Wright's accounts are at best incomplete because they fail to explain how the rules in question can be instantiated by us. Thus the rules remain unjustified. Shoemaker's account uses a rule ("Avoid Moore-paradoxical utterances") that can be motivated independently. But it is unclear how closely it is related to self-knowledge and whether it is explanatorily more basic than the phenomenon to be explained.

1. First person authority

First person authority is the authority we enjoy in self-ascriptions of certain mental states compared to ascriptions of the same types of state to other persons. We know better what we ourselves feel, believe or want

than what others feel, believe or want. Such authority is generally thought to exist in knowledge of one's own sensations and propositional attitudes, but not in knowledge of one's own emotions or character traits. The claim of "better" knowledge has at least two aspects: First, it is generally thought that we are less prone to error – some even think that we are infallible – in our self-ascriptions of sensations and propositional states. Second, it is also thought that such self-ascriptions are arrived at in a more direct, perhaps immediate way and that we do not rely on evidence or at least not on the kind of evidence that we rely on in other-ascriptions of the same mental states. Related to the second aspect is the idea that each person has such more direct access only to her own mental states and that no-one else can enjoy such access to one's own states.[1]

In communication, first person authority is manifest in the fact that we generally accept sincere assertions that self-ascribe certain mental states at face value. There seems little room to doubt or challenge them. Rather, they usually count as overriding evidence against what we might conclude from other indicators of mental states. Someone who says "I think I am going to be late" can be challenged about whether she is going to be late. It can also be doubted that her utterance is sincere. But if it is, we would normally not doubt that she really thinks so. We have what is sometimes called a "presumption of first-person authority."

2. Bar-On's neo-expressivist account

Expressivist accounts of first person authority claim that the authority is due to a special relation between the statements in which we self-ascribe mental states and the mental states themselves. According to expressivism, these statements are expressions of the mental states. What does this mean?

One inspiration for expressivism have been Wittgenstein's remarks on how we learn to use vocabulary for pain:

> [H]ow does a human being learn the meaning of the names of sensations? – of the word "pain" for example. Here is one 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-behaviour. (Wittgenstein, 1953, p. 89 [§ 244])

There are natural expressions of pain: crying, grimacing, clutching the body part that hurts etc. This pain-behaviour can be replaced by

[1] Alex Byrne has helpfully described these two aspects of authoritative self-knowledge as "privileged" and "peculiar access" to one's own mind, stressing that one does not imply the other (cf. Byrne, 2005, p. 80ff.).

exclamations such as "ouch" and, later, by sentences that we are taught by our parents. The result is a new pain-behaviour. By characterising the utterance of a sentence as "pain-behaviour," expressivists wish to indicate that the utterance holds the same kind of relation to the pain as the original natural pain-behaviour. Crying in pain is not based on the cognitive achievement of having found out something about one's inner state. Equally, it is thought, sentences that are learned as a replacement of the natural pain expression are not based on such a cognitive achievement. There is no "epistemic distance" between the utterance and the sensation because the utterance is an expression of the pain in just the same direct way as the natural pain-behaviour. And just as the natural pain-behaviour is not based on a judgment about one's inner state, neither is the self-ascriptive utterance that replaces it.

Self-ascriptive utterances with this kind of expressive character are usually called "avowals." The utterance does not express a judgment about one's pain, but rather avows it directly. Similarly, and perhaps more plausibly, expressivists claim that authoritative self-ascriptions of belief are not based on a judgment about the belief ascribed, but rather express or avow the belief directly. Likewise, self-ascriptions of intention ("I want to ϕ") do not report a judgment about the intention, an inner state, but express the intention itself. Schematically, we can say:

- The assertion "I am in pain" replaces natural pain-behaviour.
- The assertion "I want to ϕ" replaces natural "intention-behaviour." (For example, "I want the teddy" replaces reaching for the teddy.)
- The assertion "I believe that p" replaces the utterance "p."

According to expressivists, such replacement has the effect that the assertions express the states they ascribe, rather than judgments or beliefs about those states:

- The assertion "I am in pain" expresses my pain (and not my belief that I am in pain).
- The assertion "I want to ϕ" expresses my intention to ϕ (and not my belief that I want to ϕ).
- The assertion "I believe that p" expresses my belief that p (and not my belief that I believe that p).

If this is what is meant by the claim that self-ascriptions of certain mental states are expressions of those same states, how does this help to explain first person authority? Neo-expressivists such as Dorit Bar-On do not aim to account for authoritative self-knowledge but instead concentrate on our presumption that sincere assertions self-ascribing mental states are true. Bar-On's idea for an explanation of the presumption of authority is the following: We presume that such assertions are expressions of the mental

states they self-ascribe (i.e. that they are avowals in the sense in which this term has been introduced above). But to presume that a self-ascription such as "I am in mental state M" is an expression of the mental state M itself is to presume that the avowal is true. For the avowal (by me) can only be an expression of my mental state if that mental state exists (in me); and that it exists (in me), i.e. that I am in the state, is precisely what the avowal says.

> We are now in a position to offer an expressivist rendering of the presumption of truth governing avowals. To regard a linguistic act as an avowal is to take it as an expression rather than a mere report of the ascribed condition. It is to take the avowing subject to be speaking directly from her condition, where the self-ascription tells us *what* condition is to be ascribed to her. All that we as audience need to know to identify the condition being expressed is linguistic uptake. Note, however, that insofar as we take the subject to be expressing *her* condition [...], we take it that she *is* in the relevant condition – the condition that is semantically referred to by the self-ascription, which is *the very condition that would render the self-ascription true*. (Bar-On, 2004, p. 316)

So according to expressivism we presume certain self-ascriptions of mental states to be true because we presume that the subject who makes the self-ascription speaks directly from her condition, where this means that she expresses her mental state in much the same way as natural pain-behaviour expresses pain.

At this point, a clarification about the semantic status of avowals is in order. The kind of neo-expressivism defended by Bar-On recognises that avowals of mental states have as truth conditions facts about the subject's mental states. An avowal such as "I am in pain" is true if and only if I am in pain. It is just that such an avowal does not have the *role* of expressing the belief that I am in pain. Its role is to express my pain directly, not my belief that I am in pain. Likewise, an avowal such as "I believe that p" is true if and only if I believe that p. But its role is not to express my belief that I believe that p, but rather directly "to vent" my belief that p. In this neo-expressivism is different from what Bar-On calls "simple expressivism."

Simple expressivism takes it that avowals are semantically discontinuous with ascriptions from the third person. On this view, "I believe that p" *means* p and "I am in pain" does not mean anything, since it is just a different sort of pain-behaviour. As has often been noted, this view is deeply implausible because it cannot easily account for the occurrence of avowals in inferential or embedded contexts. For example "A believes that whoever believes that p is crazy" and "I believe that p" should enable me to conclude that A is inclined to believe that I am crazy. But this inference is inexplicable if "I believe that p" just means p. Similarly, statements such as "If I feel a pain in my knee I should take medicine X" do not seem to make sense if "I feel a pain in my knee" is to be analysed not as a self-ascription of pain but just as a way of clutching my knee (in pain). Simple expressivism does not seem to allow that we can make statements about our own mental states using the first person pronoun.

Bar-On's neo-expressivism avoids these difficulties by distinguishing between avowals' truth conditions and their roles. While avowals are semantically continuous with other-ascriptions, their functional role is directly to "vent" mental states rather than to express judgments about them.

The distinction between simple and neo-expressivism is important for Bar-On because it responds to an old and powerful objection against expressivism. However, in what follows I shall present a different objection which applies equally to both forms of the expressivist account. My objection is that expressivism fails to explain why we should be *capable* of reliably expressing mental states. For all the expressivist tells us there is no reason to think that avowals of mental states should be more authoritative than avowals of other, for example bodily, states self-ascriptions of which we normally do not regard as authoritative.

If the self-ascription "I am in state M" is the expression of the subject's state M, then it follows trivially that the subject is in state M. Bar-On says that we presume self-ascriptions of certain mental states to be expressions of the subject's mental states. On the basis of such a presumption, we can infer immediately and trivially that the subject is in those states. However, why do we make the initial presumption? Why do we presume self-ascriptions of certain mental states to be expressions of the subject's states but not necessarily self-ascriptions of other states, such as having low blood pressure, being of a modest character or having been born in Alabama?

Surely, Bar-On's answer has to be: because self-ascriptions of certain mental states generally *are* expressions of the subject's mental states whereas self-ascriptions of other states are not or are not equally often. But now, of course, the question arises as to what justification she can give for this claim. As we have seen, expressivists draw parallels between natural expressions of mental states (pain-behaviour, reaching for the teddy) and avowals. Bar-On also claims that avowals are "an immediate reaction to something" (Bar-On & Long, 2001, p. 326), that they "'give voice' to the subject's condition" (Bar-On & Long, 2001, p. 328), that they are "sincere, spontaneously volunteered, unreflective" (Bar-On & Long, 2001, p. 326), that we can "speak our mind" with them and so on. Jane Heal summarises the expressivist's characterisation of avowals by saying that such ascriptions are "spontaneous and in good faith" (Heal, 2001, p. 9). Now, Bar-On might be correct in describing self-ascriptions of certain mental

states in this way. But it is not clear that it follows that they should therefore invariably be regarded as expressions of the states ascribed in them.

We can see this by considering other self-ascriptions that are equally "spontaneous and in good faith" but lack special authority. For example, perhaps after some training, it might be possible to become spontaneous, unreflective and immediate in self-ascribing low blood pressure. In this sense, it might be possible directly to express one's own state of blood pressure. But however spontaneous and immediate the utterance of "I am having such a low blood pressure," it seems unlikely that it should ever have as much authority as the self-ascriptions of, say, beliefs or intentions (cf. Heal, 2001, p. 9). When successful, the self-ascription of low blood pressure here seems to be an expression of this state. But its success rate might be rather low.

Another example, given by Alex Byrne, is the self-ascription of present perceptual states. The assertion "I see a red cardinal" can be just as spontaneous and immediate as an authoritative avowal of belief (cf. Byrne, 2011, p. 716). Again, it seems that if true, such an assertion could be regarded as a direct expression of the subject's perceptual state. But, of course, it is clear that even a spontaneous, unreflective and immediate self-ascription of a perceptual state can easily go wrong. Perception might deceive us or our classificatory skills might fail us. Certainly the spontaneity and unreflectiveness of our assertion is no reason to ascribe a higher degree of authority to it.

The case of proprioception is similar. It seems that I can avow immediately that I have my legs crossed. Everything Bar-On says about avowals of mental states seems to be true of such an avowal of a bodily state as well, except for a comparable lack of authority. Proprioception can go wrong and probably more easily so than self-ascriptions of mental states. So again, a high degree of immediacy, spontaneity, unreflectiveness and so on does not guarantee that the self-ascription in question is an expression of the state ascribed.

These examples show that the expressive character of true avowals of mental states is not sufficient to explain their authority. True self-ascriptions of present perceptual states or of certain bodily states can be equally expressive in character. But ascriptions of this kind can also easily be false and so fail to be expressive of the state that the assertion ascribes to the subject. *If* a self-ascription of some state to the subject is an expression of this same state in the subject, *then* it is true. But when are self-ascriptions expressive in this way? What reason is there to think that self-ascriptions of mental states generally are expressive in this way, while self-ascriptions of other states are not? Why is it that, as Bar-On says, "I can speak my mind, but cannot speak my body" (Bar-On, 2004, p. 428)? It

seems that the expressivist account of first person authority is at best incomplete as long as it does not answer this question and Bar-On has surprisingly little to say about it.[2]

One way to formulate this problem is as follows. Expressivists such as Bar-On correctly identify a language structure that guarantees truth in certain self-ascriptions. A self-ascription is necessarily true if it is an expression of the same state that it ascribes to the subject. Expressivists claim that the authority of self-ascriptions of mental states is due to the fact that they instantiate such a linguistic structure. Now the problem is that the expressivists do not tell us how it is possible for us to instantiate this language structure. We might put it thus: We are given rules as to how mental predicates are to be interpreted in the case of self-ascriptions. They are generally to be seen as expressions of the states they describe. But given these rules, it is unclear why anyone should be capable of acting so as to be interpretable according to these rules; especially so, since other predicates, for example those ascribing perceptual states, clearly cannot be used by us in a way that would allow us to be interpreted according to similar rules. It is not sufficient for an explanation of first person authority to show that there are rules for the interpretation of mental predicates which imply the truth of mental self-ascriptions. Rather, it must also be shown how it is possible for us to have such rules for mental predicates given that we clearly cannot have them for other ones.

3. Wright's constitutive account

A variety of constitutivist accounts of first person authority have been proposed in the recent debate (e.g. Shoemaker, 1990; Bilgrami, 1998; Stoneham, 1998; Heal, 2001; Moran, 2001). Here I shall briefly discuss the proposal by Crispin Wright[3] because his way of relating rules of language and first person authority seems to me to be vulnerable to an objection similar to the one put forward against Bar-On's expressivism.

Just as expressivism, Wright's account is also inspired by Wittgenstein. But it does not attempt to infer the authority from some similarity between natural expressions and mental self-ascriptions. And unlike simple expressivism, it does not deny the ascriptive character of avowals.

---

[2] She affirms that neither perceptual or proprioceptive states nor "purely physical conditions, such as having a cold or a diseased state of one's liver" (Bar-On & Long, 2001, p. 330) can be expressed in the same sense as mental states. But these affirmations look like "terminological stipulation" (Byrne, 2011, p. 716) since self-ascriptions of such states often seem equally immediate, spontaneous and in good faith as self-ascriptions of mental states.

[3] As put forward in Wright (1989a). Other relevant texts include Wright (1989b) and his Whitehead Lectures (=Essays 10 and 11 in Wright, 2001).

According to Wright, first person authority is due to the constitutive relation between self-ascriptions of mental states and those states themselves. The important point is not that such self-ascriptions are expressions of the states ascribed in them. (Wright is silent on whether or not they are.[4]) Rather, it is the fact that they are criterially related to them. The self-ascription of a mental state is a defeasible criterion for being in that state. It is neither based on an observation nor on an inference, because it is not a contingent by-product of the state. The relation between the self-ascription and the state ascribed is not causal and contingent; rather, it is a priori. Self-ascribing a mental state is part of what it means to be in that state.

On this account, self-ascriptions of mental states are true by default. It is not the case that they can never be false. It might be that under certain circumstances the rest of what a person says and does make it plausible that she is not in the state that she self-ascribes. But under normal conditions the self-ascription of a mental state has to be taken as true because its occurrence constitutes the state ascribed. The self-ascription is extension-determining, not extension-reflecting.[5]

Similarly to Bar-On, Wright focuses on the question of why we presume that others' self-ascriptions of mental states are authoritative. His answer seems to be that this is something like a primitive given of our language game, not capable of further explanation. It is simply a feature of the "grammar" of mental self-ascriptions. These are not based on some kind of cognitive access to our own mind, which could perhaps break down in the way perception or inferences can be misguided. Rather, they are normally true simply because our language game requires us to presume them to be true. The authority is thus akin to a concession by the other participants in the language game, not to a special cognitive achievement.

In a sense, then, Wright's account is even thinner than Bar-On's with regard to the mechanism by which we manage to be authoritative in our mental self-ascriptions. Bar-On appeals to the similarity with natural expressions (and I have argued that this does not suffice to explain the authority). Wright simply declares the authority to be a feature of our language game.

[4] In Wright, 1989a. In his Whitehead Lectures, Wright shows some sympathy for the expressivist proposal, saying that it "flies rather further than is usually thought. But it is a dead duck all the same" (Wright, 2001, p. 364). He finds that it does well in explaining the authority of avowals, but fails to account for unexpressed, yet authoritative self-knowledge (cf. 2001, p. 363f.).

[5] The terms "extension-determining" and "extension-reflecting," as applied to judgments in general as well as avowals in particular, is from Wright, 1989b, p. 192ff.

However, he is admirably clear about the presuppositions of his account. He declares that it presupposes "certain deep contingencies" (Wright, 1989, p. 632): It must be the case that taking others' self-ascriptions of mental states to be authoritative indeed enables us to understand them better than not making this assumption. Furthermore, self-ascribers must make such true self-ascriptions "just like that," i.e. not because they recognise them to be true (a cognitive achievement), but because the ascriptions simply "come to them" in the right moment.

> Since the telos, in the most general terms, of the practice of ascribing intentional states to oneself and others is mutual understanding, the success of a language game that worked this way would depend on certain deep contingencies. It would depend, for instance, on the contingency that taking the self-conceptions of others seriously, in the sense involved in crediting their beliefs about their intentional states, as expressed in their avowals, with authority, will almost always tend to result in an overall picture of their psychology which is more illuminating – as it happens, enormously more illuminating – than anything which might be gleaned by respecting all the data except the subject's self-testimony. And that in turn rests on the contingency that we are, each of us, ceaselessly but – on the proposed conception – subcognitively moved to opinions concerning our own intentional states which will indeed give good service to others in their attempt to understand us. Thus, we do not cognitively interact with states of affairs which confer truth upon our opinions concerning our own intentional states; rather, we are inundated, day b[y] day, with opinions for which truth is the default position, as it were. (Wright, 1989, p. 632f.; cf. also Wright, 2001, p. 313)

Wright's proposal has been widely discussed and criticised in various points. It might be charged, for example, that he cannot account for the idea that we have genuine *knowledge* of our mental states (Fricke, 2008, p. 78f.). It could also be argued that his account makes it mysterious how mental states, constituted just by my opinion that I have them, can be causally efficacious (Heal, 2001, p. 16). Other questions are: Is it not possible for someone's self-ascription to confirm, or be confirmed by, what we otherwise know about the person? Wright's account seems to deny this because he seems to suggest that, under normal conditions, the self-ascription makes itself true (Smith, 1998, p. 413). – Further, since we sometimes seem to have first-order mental states (such as beliefs) without knowing that we do, why is it that *only some* first-order mental states are constituted by self-ascriptions of those states, while others, perhaps with the same content, are not (Smith, 1998, p. 413f.)?

Here I shall only make a critical point which is analogous to the one made earlier about Bar-On's expressivism and which deserves attention because it highlights a general difficulty for theories that aim to explain first person authority by reference to language structures. Suppose we accept Wright's claim that the relation between self-ascriptions and mental

states is a priori and not of an epistemological nature. The question must arise then as to how it is possible for us to make self-ascriptions about our own mental states.

The reason this question must arise is the following: Clearly, there are many predicate-ascriptions, to ourselves or other persons or objects, that are not constitutively related to the states they describe. Rather, they are contingently related to them. But of course we can imagine a language game in which some such descriptions are not contingently, but equally constitutively related to their subject matter. They would then enjoy the same sort of authority as our mental self-ascriptions. For example, the statement "Tomorrow it will rain" could be criterially related to tomorrow's rain. The relation between rain and statement would be a priori; under normal conditions, the statement would determine that it will rain, not reflect the fact. Uttering "Tomorrow it will rain" would not be based on a cognitive achievement, but be an opinion to which I am subcognitively moved and which would give good service to others in determining whether it is going to rain the next day. The structure of this language game would be analogous to the one Wright describes about self-ascriptions of intentional states.

Evidently, we are not able to play such a language game. Why not? The relevant deep contingencies are missing. We are not, generally, subcognitively moved to opinions about tomorrow's rain. And when we are, such opinions are not a good guide to tomorrow's weather. The only way for us to determine whether it is going to rain is through observation and meteorology. The degree of reliability in our weather forecasts can be explained by facts about the weather, the richness of our data and our insight (or lack of it) into the laws of meteorology. These are important contingent facts about the relation between the rain and the statement about it.

Let us compare this case with that of mental self-ascriptions. We can authoritatively self-ascribe mental states. So we can, perhaps, play a language game of the kind Wright describes. But what, exactly, is the difference between the self-ascriptions game and the rain game? It seems that the important difference lies in the "deep contingencies." It is because of the way we are made (and related to rain and to our mental states) that we can play a language game in which self-ascriptions of mental states seem to be constitutively related to the states ascribed, but cannot play one in which forecasts of rain are constitutively related to coming rain. In other words, what explains the authority of mental self-ascriptions has to do with what Wright describes as the "deep contingencies" of the respective language game. The structure of the language game itself (that one element seems to be constitutively related to another) apparently has very

little explanatory value when it comes to accounting for first person authority.

This result is similar to what we found in the case of Bar-On's theory. Wright describes the rules of a language game that guarantees default truth to self-ascriptions of mental states. Such self-ascriptions have to be regarded as constitutive of their own truth. However, even if we accept this characterisation of the way we understand ascriptions of mental states, the theory leaves open the question as to why it is the case that we are capable of playing this language game. Why is it the case that we can instantiate a language game of the kind Wright describes but not an analogous language game in which rain forecasts are constitutive for future rain? Wright's theory does not answer this question. It only points to "deep contingencies" that we have to assume as underpinnings of the language game he describes. But it seems that these contingencies do the real work in his account. How is it possible that we generally have true opinions about our own mental states but not necessarily about other persons' states (or about the future rain)? It seems that a theory of first person authority should attempt to answer this question, rather than just stating that this is so.[6]

[6] It might be said that my argument begs the question of whether it makes sense to demand of philosophy that it provide an explanation of features of our language games. Wright repeatedly points out that Wittgenstein favoured a "descriptive method" in philosophy: "We must do away with all *explanation* and description alone must take its place." (Wittgenstein, 1953, p. 47 [§ 109]) "Our mistake is to look for an explanation while we ought to look at what happens as a 'proto-phenomenon'. That is, where we ought to have said: *this language game is played.*" (Wittgenstein, 1953, p. 167 [§ 654], cf. Wright, 2001, p. 364f.) On Wright's interpretation, it is because of this conception of philosophy that Wittgenstein does not offer a more substantive account of the authority we enjoy in mental self-ascriptions. Wright notes that this position "can seem intensely unsatisfying" (2001, p. 317), but ultimately he seems to endorse it. It is not entirely clear what reasons can be given in its favour (partly because reason-giving would seem to be a form of philosophical explanation), except, perhaps, for the demonstration that all alternative proposals have serious shortcomings (cf. Wright, 2001, p. 373). – I cannot here engage in any depth with Wittgenstein's conception of philosophy. I shall only remark that it does seem to be a substantive question what kind of empirical underpinnings are necessary for a given language game to be realised by us. This is why Wright himself articulates his "deep contingencies." Furthermore, it seems evident that such underpinnings can be further investigated and illuminated through scientific inquiry. In our case, the science in question might be cognitive science. Perhaps the participation in such inquiry should not be regarded as "proper" philosophy. But philosophers do participate fruitfully in such inquiry, evaluating coherence and plausibility of theories and proposing new hypotheses. I am happy to characterise this paper as an exercise in such "improper" philosophy.

4. Moore's paradox and self-knowledge (Shoemaker)

In "On Knowing One's Own Mind" (1988), Sydney Shoemaker relates knowledge of one's own beliefs with the capacity to recognise the awkwardness of uttering Moore-paradoxical sentences. It seems to me that this is the most promising way of explaining first person authority by reference to the rules of language.

Shoemaker's argument has the form of a *reductio ad absurdum*. Suppose that George is self-blind. He has no direct, first-person access to his own beliefs. The only way he can know about his beliefs is in the way we can also know about the beliefs of other persons. Suppose, however, that otherwise George has normal intelligence and normal cognitive and conceptual abilities. He can have beliefs about the world and beliefs about other people's beliefs. He can also understand and have beliefs about his own beliefs. His self-blindness just means that he cannot access them in a direct, first-person way. Shoemaker's argument aims to establish that such a person should be capable of recognising the paradoxical character of Moore-sentences and that this should enable him to self-ascribe beliefs in just the same direct and exclusively first-personal way as we do. Hence self-blindness is impossible in persons with normal cognitive and conceptual abilities.

Moore-paradoxical sentences are sentences of the form "p, but I don't believe that p." Clearly, the proposition expressed by such a sentence can be true. Perhaps most truths (but if not most, at least very many) obtain without being believed by me. Yet, although the proposition expressed by a Moore-paradoxical sentence can be true, it seems that such a sentence cannot be coherently asserted. The explanation Shoemaker gives is pragmatic. Assertions, if sincere, express beliefs. So the first conjunct of Moore's sentence, if asserted sincerely, expresses the subject's belief that p. However, the second conjunct says that the subject does not have this belief. So the first conjunct expresses a belief whose existence refutes what the second conjunct says. This means that "one could not hope to get one's audience to accept both conjuncts on one's say so, and could have little hope of getting them to accept either" (Shoemaker, 1988, p. 194). More precisely, one could not hope to get one's audience to accept the first conjunct as an expression of one's belief and the second conjunct as true. But it is the pragmatic purpose of assertions to be taken both as true and as expressions of the utterer's beliefs. Since Moore-paradoxical sentences cannot be taken both as true and as expressions of the subject's belief, asserting them is self-defeating.

Since George has normal cognitive and conceptual capacities he should be able to recognise the pragmatically self-defeating character of Moore-paradoxical sentences. When hearing someone utter such a sentence he

should be just as perplexed as we are. This means that he should also refrain from uttering such sentences himself. But avoiding Moore-paradoxical sentences while still making assertions about the world can lead directly to self-ascriptions of belief. The reasoning could proceed along the following steps:

(1) George knows that p.
(2) He knows that he must not utter Moore-paradoxical sentences. (He knows that they are pragmatically self-defeating.)
(3) Now he can infer that he must not deny that he believes that p.
(4) This means that he believes that p.

If someone asks George whether he believes that p, he can answer by asking himself whether p and inferring that he believes that p in case he finds that p. This procedure is exclusively first-personal. One cannot find out whether someone else believes that p by asking oneself whether p. Probably there are also good reasons to believe that the procedure is especially reliable. It just involves a simple inference, no perception and no complex reasoning. So it seems that George has, after all, an exclusively first-personal and even especially reliable access to his own beliefs. Shoemaker claims that George would be indistinguishable from us. So the claim that there could be a person who is self-blind, but has normal intelligence and normal cognitive and conceptual abilities has been reduced to absurdity.[7]

In the context of the previous discussion, we can put this result as follows: Shoemaker shows that if a normal language speaker follows the rule not to utter Moore-paradoxical sentences, then she can by way of very simple inferences come to make authoritative self-ascriptions of belief. Shoemaker even shows why such a rule is pragmatically necessary. As far as I have sketched it, this explanation of first person authority only applies to self-ascriptions of belief.[8] But it seems to me that it is far more

[7] This conclusion might be too fast. Shoemaker does not show that we in fact proceed in the same way as George. So perhaps George just has a specially developed capacity for self-ascribing beliefs from a third-personal access, while we ascribe them in some other exclusive way. For an objection of this type see Kind (2003, p. 45ff.) and Byrne, who puts it this way: "Why hasn't Shoemaker just outlined a strategy for *faking* self-knowledge?" (Byrne, 2005, p. 92). To avoid this kind of objection we should read Shoemaker as giving an hypothesis as to how we might, in fact, come to acquire knowledge of our own minds.

[8] Shoemaker tries to extend his account to knowledge of desires (and hints at how it might work with states such as hope and intention). His starting point is the claim that in the case of these states there are counterparts to Moore's paradox such as "Please close the window, but I don't want you to," "Would that he would

promising than the attempts by Bar-On and Wright to explain the authority by reference to language structures. There is no mystery here of how it is possible for us to follow the rule that produces first person authority. Normal linguistic abilities are enough both to recognise that Moore-paradoxical sentences are self-defeating and to be disposed not to utter them. Normal cognitive abilities suffice then to make the simple inferences that produce authoritative self-ascriptions of belief.

This Moore-inspired account of first person authority is very close to contemporary "transparency" theories of self-knowledge. These take their cue from Gareth Evans's famous remark that "I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*" (Evans, 1982, p. 225). Alex Byrne has encapsulated the procedure neatly in the following rule:

BEL If *p*, believe that you believe that *p* (Byrne, 2005, p. 95)

Evans can be naturally interpreted as suggesting that by following BEL we can make authoritative self-ascriptions. Shoemaker's account from 1988, in turn, can be seen as providing an independent justification for rules such as BEL. On this view, it is because we have to avoid Moore-paradoxical utterances (and we know that he have to) that we self-ascribe beliefs by following BEL. So considerations about the rules of language, more specifically about the pragmatics of Moore-paradoxical sentences, help to give an independent justification for an epistemic account of self-knowledge. BEL is not only a good rule because, as Byrne says, it is self-verifying (even merely trying to follow it but getting one's facts about the world wrong and starting from a falsehood p, produces true self-ascriptions of belief) and because it just requires very basic cognitive capacities (no observation or complex inferences). It is also a good rule because it enables us to avoid Moore-paradoxical sentences, as every competent language user must strive to if he wishes to be understood.

However, I shall conclude my discussion with two critical remarks about the Moore-inspired account of first person authority. First, the reasoning from avoidance of Moore-paradoxical utterances to authoritative self-ascriptions is not as straightforward as I have suggested above. Given the pragmatic account of the paradoxical character of Moore's sentences, the subject's reasoning should probably go as follows:

(5) p
(6) I must not assert "p, but I don't believe that p."
(7) I must not assert "I don't believe that p."

come, but I hope that he doesn't," and "I'll be there, but I intend not to be" (cf. Shoemaker, 1988, p. 204ff.).

(8) I must not deny "I believe that p"
(9) I believe that p.

The inference from (5) to (9), without the intermediate steps, corresponds to Byrne's rule BEL. (6) (7) and (8) express the additional justification provided by the pragmatic account of Moore's paradox. But do they really help to justify the conclusion? One problem with this suggestion is that (6) to (8) are not statements about my beliefs or about p. Rather, they are statements about what I should *assert*. And the reasons for the recommendations in (6) to (8) are supposed to be pragmatic. Given that p and given that, therefore, I should assert that p, I should not deny that I believe that p. But of course the conclusion (9) is not supposed to be that I should *say* that I believe that p. It is supposed be that it is *true* that I believe that p, a piece of self-knowledge. It seems that pragmatic reasons cannot really support that conclusion.

The situation would be different if Moorean sentences were straightforwardly contradictory. It then simply could not be *true* that p but that I do not believe that p. In this case, given p, I could directly infer that I believe that p. But one defining characteristic of Moorean sentences is that they *can* be true. The paradox is precisely that they are in some sense impossible despite possibly being true. Now, if the impossibility is seen as merely pragmatic, then it is hard to see how it can support any conclusion about what I really do believe – as opposed to what I should or should not *assert* about what I believe.

My second critical remark concerns the question of explanatory basicness. Even if considerations about Moore's paradox implied the correctness of the Evansian procedure as summarised by BEL, this does not mean that they can explain and justify it. The explanatory relation might, rather, be the reverse. Perhaps the paradoxical nature of Moorean sentences is explained by the self-knowledge we can acquire through BEL. Or both Moore's paradox and BEL can be explained by an independent account of self-knowledge. Shoemaker himself suggests in a later article (Shoemaker, 1995) that considerations about self-knowledge explain the paradoxical character of Moorean sentences and not vice versa. He cites as evidence for regarding self-knowledge as explanatorily more basic the fact that apparently Moorean sentences do not have to be asserted to be paradoxical. It seems equally impossible merely to *believe* the content of a Moorean sentence. If this is true, if it is incoherent to believe "p, but I don't believe that p" whether or not the belief is also expressed in an assertion, then we might have to explain the paradox without reference to the pragmatic conditions of assertion. This is not a necessary conclusion, but it is at least plausible.

I shall not examine Shoemaker's own account of the paradoxical nature of Moore-sentences, since it leads us to a different sort of constitutive

account of self-knowledge.[9] Here it should suffice to point out that if BEL can provide us with knowledge of our own beliefs, then it can also serve to explain at least part of Moore's paradox. Someone who believes that p and has BEL at her disposal will also be inclined to believe that she believes that p. But this piece of self-knowledge is in contradiction with the second conjunct of the Moorean sentence. Believing "p, but I don't believe that p," she will also be disposed to believe that she believes that p (by applying BEL to the first conjunct of the Moorean belief). And this means that she will be disposed to have contradictory beliefs, namely "I believe that p" and "I don't believe that p." This would explain why Moorean sentences are not only self-defeating assertions but also awkward when believed.

There are independent reasons for regarding BEL as a reliable rule we actually use – its self-verifying character; the fact that only ordinary simple cognitive capacities are necessary to apply it; and the fact that it corresponds to the observation that when we are asked about our beliefs we think about the world rather than about our inner states. It might also be plausible to regard considerations about belief and knowledge as more basic than considerations about assertability. All this would support the conclusion that BEL is explanatorily more basic than the pragmatic impossibility of Moorean sentences.

We have seen that the Moore-inspired account looked much more promising in its attempt to use rules of language to explain first person authority. It has no problems with explaining how it is possible with ordinary cognitive capacities to follow the rules that do the explanatory work and it has an independent justification for these rules. It seemed that the account provided independent, language-based support for epistemic transparency theories of first person authority of the kind suggested by Evans and Byrne. However, on closer inspection, it seems that the proposed pragmatic elucidation of Moore's paradox is less closely related to self-knowledge than the account suggests and that the paradox might actually have to be explained with the help of a theory of self-knowledge rather than vice versa.

[9] Shoemaker argues that the capacity to rationally adjust one's belief-desire system in the face of new information requires one to know which beliefs and desires one has. So the constitutive relation is between rationality and (privileged) self-knowledge. Most of his arguments can be found in Shoemaker (1988) and (1990).

5. Conclusion

In this paper I have looked at three different proposals to use rules of language to explain first person authority. Dorit Bar-On's neo-expressivist proposal and Wright's constitutivist account seem to suffer from similar problems: The structures of language they describe rather trivially entail the presumption of first person authority. But the proposals fail to explain how it is possible for us to instantiate these structures of language or how it is possible to follow the linguistic rules they identify in our language. Thus the accounts remain unsatisfying; at best incomplete and possibly vacuous. The third account, suggested by Sydney Shoemaker, is based on the uncontroversial claim that anyone with ordinary cognitive and conceptual abilities should recognise the paradoxical nature of Moore-sentences. It attempts to derive from this fact a simple way of acquiring self-knowledge of one's own beliefs which coincides with contemporary transparency theories of the mind, inspired by Evans. If this third account were successful, it would provide an interesting, independent justification for the transparency theories. Alas, the relation it postulates between a pragmatic ability to recognise the paradoxical nature of Moore-sentences and self-knowledge can be questioned and there are reasons to regard a correct account of self-knowedge as explanatorily more basic than the correct theory of Moore's paradox. It might be useful, then, to explore epistemic or other accounts of first person authority before trying to relate the phenomenon to our language.[10]

References

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

Bar-On, D. & Long, D.C. (2001). Avowals and First-Person Privilege. *Philosophy and Phenomenological Research, 62* (2), 311-335.

Bilgrami, A. (1998). Self-Knowledge and Resentment. In Wright, C., Smith, B., & Macdonald, C. (Eds.), *Knowing Our Own Minds* (207-241). Oxford: Oxford University Press.

[10] The main ideas of this paper were presented at the congress "Semantics and Philosophy in Europe 4" which took place in Bochum, Germany, in 2011. I am grateful for the comments of the audience on that occasion. I wrote the paper while on sabbatical leave at Ruhr-Universität Bochum, which was generously funded by the National Autonomous University of Mexico through its Programa de Apoyos para la Superación del Personal Académico de la UNAM. Finally, I would also like to thank an anonymous referee for the *Polish Journal of Philosophy* for some interesting comments.

Byrne, A. (2005). Introspection. *Philosophical Topics, 33* (1), 79-104.
Byrne, A. (2011). Review Essay of Dorit Bar-On's *Speaking my Mind. Philosophy and Phenomenological Research, 83* (3), 705-717.
Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.
Fricke, M. F. (2008). Teorías constitutivas de la autoridad de la primera persona: Wright y Heal. *Ludus Vitalis, 16* (29), 73-91.
Heal, J. (2001). On First Person Authority. *Proceedings of the Aristotelian Society, 102* (1), 1-19.
Moran, R. (2001). *Authority and Estrangement*. Princeton: Princeton University Press.
Shoemaker, S. (1988). On Knowing One's Own Mind. *Philosophical Perspectives, 2*, 183-209.
Shoemaker, S. (1990). First-Person Access. *Philosophical Perspectives 4*, 187-214.
Shoemaker, S. (1995). Moore's Paradox and Self-Knowledge. *Philosophical Studies, 77* (2-3), 211-228.
Smith, B. (1998). On Knowing One's Own Language. In Wright, C. & Smith, B. & Macdonald, C. (Eds.), *Knowing Our Own Minds* (391-428). Oxford: Oxford University Press.
Stoneham, T. (1998). On Believing that I am Thinking. *Proceedings of the Aristotelian Society 98* (2), 125-144.
Wittgenstein, L. (1953). *Philosophical Investigations*. Ed. by G.E.M. Anscombe and R. Rhees. Translated by G.E.M. Anscombe. Oxford: Blackwell.
Wright, C. (1989a). Wittgenstein's Later Philosophy of Mind. Sensation, Privacy and Intention, *Journal of Philosophy, 86* (11), 622-634.
Wright, C. (1989b). Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics. In George, A. (ed.), *Reflections on Chomsky* (233-264). Oxford: Blackwell. (Reprinted in Wright, 2001, pp. 170-213. Page numbers are from the reprint.)
Wright, C. (2001). *Rails to Infinity. Essays on Themes from Wittgenstein's* Philosophical Investigations. Cambridge, Mass.: Harvard University Press.

POLISH JOURNAL OF PHILOSOPHY
Vol. VI, No. 2 (Fall 2012), 33-52.

Blurring: Structural Realism and the Wigner Puzzle

Alexander James Gillett
University of the West of England

Abstract. Investigating the metaphysical problem of nature requires engaging with philosophy of science. Arguments in this field, combined with metaphysical underdetermination problems in fundamental physics, have given rise to a sophisticated form of scientific realism called ontic structural realism; and the re-conceptualisation of metaphysics in terms of structures. This transforms the problem of nature into the dissolution of the distinction between mathematical and physical structures (what we shall call the "blurring problem"). To date, there has been an insufficient exploration of this problem in the literature because it has been deemed unscientific. This essay demonstrates that the problem is legitimate, important, and connects with a wider issue in the philosophy of mathematics—namely, the problem of applicability of mathematics to the sciences' investigation of nature (the Wigner Puzzle).

1. Introduction

This essay examines how, through a necessary engagement with philosophy of science, the metaphysical problem of nature becomes transformed into an examination of the "dissolution" of the distinction between mathematical and physical structures. Ontic structural realism urges us, on the basis of findings from fundamental physics and the history of science, to abandon metaphysics composed of individuals and self-subsistent objects (object-orientated realism) because it is both an inadequate ontology for fundamental physics and also belies an anthropocentric bias that a true realism should surpass. Instead, ontic structural realism proposes a metaphysics of structures. This has a radical outcome for the metaphysical problem of nature, which consequently becomes suffused with the question "what is structure?" and re-conceptualised as *the problem of the dissolution of the distinction between the mathematical and the physical*—or, more, incisively: what is the relationship between mathematical and physical structures? Henceforth, for the sake of brevity, we shall refer to this as the "blurring problem." Properly examining this question—a task omitted from the structural realist literature to date—will show the links between the blurring problem and an issue in the philosophy of mathematics: the