Routledge
Taylor & Francis Group

Philosophical Explorations

An International Journal for the Philosophy of Mind and Action

ISSN: 1386-9795 (Print) 1741-5918 (Online) Journal homepage: http://www.tandfonline.com/loi/rpex20

Transparency, expression, and self-knowledge

Dorit Bar-On

To cite this article: Dorit Bar-On (2015) Transparency, expression, and self-knowledge, Philosophical Explorations, 18:2, 134-152, DOI: 10.1080/13869795.2015.1032334

To link to this article: http://dx.doi.org/10.1080/13869795.2015.1032334

Philosophical Explorations, 2015
Vol. 18, No. 2, 134–152, http://dx.doi.org/10.1080/13869795.2015.1032334
This paper is part of the special issue 'Self-knowledge in perspective'
guest edited by Fleur Jongepier and Derek Strijbos

Transparency, expression, and self-knowledge

Dorit Bar-On*

Philosophy Department, University of Connecticut, Storrs, CT, USA

(*Received 10 May 2014; final version received 10 March 2015*)

Contemporary discussions of self-knowledge share a presupposition to the effect that the only way to vindicate so-called first-person authority as understood by our folk-psychology is to identify specific "good-making" epistemic features that render our self-ascriptions of mental states ('avowals') especially knowledgeable. In earlier work, I rejected this presupposition. I proposed that we separate two questions:

(i) How is first-person authority to be explained?
(ii) What renders avowals instances of a privileged kind of *knowledge?*

In response to question (i), I offered a *neo-expressivist* account that, I argued, is compatible with a variety of non-deflationary, substantive answers to question (ii). Here I re-evaluate the relative merits of the neo-expressivist account in light of some recent attempts to capture first-person authority by appealing to the so-called *transparency* of mental self-attributions. I then canvass two recent appeals to transparency that give priority to question (ii). Bearing in mind difficulties with the recent attempts, I return (in Section 3) to relevant aspects of my own neo-expressivist account, which, instead, begins with question (i). I conclude by offering reasons for thinking that the neo-expressivist approach is better suited for integration with a folk-psychologically grounded understanding of ourselves than the alternatives canvassed.

Keywords: self-knowledge; transparency; expression; first-person authority; neo-expressivism; folk psychology; avowals

1. Introduction

Our daily commerce with fellow human beings is rife with spontaneous pronouncements about our current states of mind ("I have a terrible headache", "I'm hoping Joe gets here on time"). And we regularly solicit such pronouncements from others ("Hungry?", "Want some coffee?", "Would you agree that was a good film?" and so on). As common sense would have it, our spontaneous or solicited present-tense mental self-ascriptions – avowals, as I shall refer to them – enjoy a kind of *epistemic security* unparalleled by any other reports concerning contingent matters of fact (including many self-reports). We ordinarily take others' avowals at face value; we strongly presume them to be true, we do not suppose it appropriate to challenge or correct them (*qua self*-attributions), and we regard our own avowals as a place of "epistemic retreat", in the sense that we do not take them as standing in any need of justification or reasons. There is also a strong presumption in everyday discourse, and even in contexts of therapy or psychological research, that our "unstudied utterances" (as Ryle referred to them) constitute a kind of psychological benchmark.[1]

*Email: dorit.bar-on@uconn.edu

But philosophers and psychologists famously disagree on how best to explain this presumption of so-called *first-person authority*. It seems clear to me that it cannot be plausibly thought to be *simply* a matter of a courtesy we pay our fellow human beings. At the same time, the status accorded avowals by folk psychology can seem especially puzzling – downright baffling – if we approach it with very specific epistemic theoretical commitments. For while avowals do not seem to be at all like (what Sellarsians would describe as) "moves in the game of asking and giving reasons" – statements that require and for which we can provide justification – it seems wrong to suggest that (therefore) they are simply not items of knowledge at all (as some Wittgensteinians would have it).

In Sellars' famous myth of Jones, we find a rather different approach to explaining the epistemic status of avowals. Sellars provides a particular rational reconstruction of the *origins* of the commonsense conception of "mental acts construed as *pure* occurrents". He envisages a genius Jones who "develops a *theory* according to which overt utterances are but the culmination of a process which begins with certain inner episodes" that can be understood as functionally equivalent to meaningful utterances spoken out loud.[2] Following Jones, our "Rylean ancestors" might have adopted a *theory* "designed to *explain* [behavioral] propensities to think-out-loud", on the model of a micro-physical theory that is designed to explain the behavior of perceived macro objects (Sellars 1975, Lec. II, Sec. 52). As for the commonsense ideas of privileged access and of non-inferential self-knowledge, Sellars suggests (somewhat cryptically) that, once Jones' Rylean fellows learn to explain each other's behavior by appeal to covert inner episodes (and given Rylean capacities already in place), it is a "short step" for Jones and his fellows to move from what was initially "a language with a purely theoretical use" to self-attributions that have a self-reportive, non-inferential (and in that sense privileged) "avowal role".[3]

Like many contemporary empiricist introspectionists, Sellars emphasizes that privileged access, as he explains it, entails reliability but *not* infallibility, though Sellars does not appeal to any perception-like introspective *mechanism* or *route*. And, unlike contemporary rationalists, Sellars does not invoke the *self-intimating* character of mental states, even though his account *is* offered as part of a broader account of the *nature* of mental states that aims to do justice to both normative and naturalistic aspects of mind. Yet, unlike philosophers in both camps, Sellars does *not* begin his investigation with the supposition that there is a special puzzle about how our mental self-attributions could possibly qualify as knowledge. His primary focus is on the folk-psychological character and *use* – what he calls the "avowal role" – acquired by certain self-reports.[4]

I see Sellars' explanation as representing a departure from a certain presupposition that is rarely seriously questioned by parties to the discussion of self-knowledge. The presupposition is to the effect that the only way to vindicate first-person authority as understood by our folk psychology is to identify specific "good-making" epistemic features that render avowals especially knowledgeable.[5] In earlier work, I rejected this presupposition. I proposed that we separate two questions:

(i) What could vindicate the commonsense notion of first-person authority – or, how is avowals' distinctive security to be explained?
(ii) What renders avowals instances of a privileged kind of *knowledge*?

In response to question (i), I offered a *neo-expressivist* account that, I argued, unlike more familiar expressivist accounts, is compatible with a variety of non-deflationary, substantive answers to question (ii).[6] In what follows, I wish to reevaluate the relative merits of the neo-expressivist account in light of some recent attempts to capture first-person authority by

appealing to the so-called *transparency* of mental self-attributions. In the next section, I will canvass two recent appeals to transparency that give priority to question (ii). Bearing in mind difficulties with the recent attempts, I review – in Section 3 – relevant aspects of my own neo-expressivist account, which, instead, begins with question (i). In Section 4, I offer some reasons for thinking that the neo-expressivist approach is better suited for integration with a folk-psychologically-grounded understanding of ourselves than the alternatives canvased.

2. Transparency[7]

The fundamental puzzle about self-knowledge seems to be this. Our ordinary avowals appear at once *epistemically groundless*, in the sense of not requiring or relying on any reasons or justification, and *epistemically privileged*, in the sense of representing a distinctive kind of *knowledge* of our states of mind.[8] If taken at face value, avowals' groundlessness may incline us toward *deflationism* about self-knowledge – the view that, insofar as avowals lie outside the epistemic "game" of justification and reasons, they do not represent genuinely knowledgeable claims or beliefs *about* our states of mind. As against deflationism, a number of authors have in recent years denied the appearance of avowals' groundlessness and tried to explain it by recalling a thought found in Evans (1982) (and later developed in a number of different ways by several authors) (see, e.g. Gallois 1996; Moran 2001; Fernández 2003; Byrne 2005). This is the idea that when producing self-attributions of current states of mind, such as "I believe that it's raining", "I'm having a visual sensation of something blue", we do not normally attend inwardly, as it were, to the contents of our mind (as introspectionist and acquaintance views would have it).[9] Instead, we attend to the same *outward* facts, objects, or properties that we would attend to if we were considering, for example, whether it is raining, or whether there is something blue in front of us. When making such self-attributions, we look *through* the attributions to the worldly features at which they are directed.[10] Call this "transparency-to-the-world".[11]

Evans introduces transparency-to-the-world (though not under that label) as part of his attempt to avoid the Cartesian conception of "self-knowledge as a form of perception – mysterious in being incapable of delivering inaccurate results" (1982, 225). In a much-cited passage, Evans observes that, if asked: "Do you think there is going to be a third world war?", I must "attend to precisely the same outward phenomena as I would attend to if I were answering the question '*Will* there be a third world war?'". He goes on to identify a "procedure for answering questions about what one believes" that can be encapsulated "in the following simple rule: whenever you are in a position to assert that *p*, you are *ipso facto* in a position to assert 'I believe that *p*'" (Evans 1982, 225). As Evans remarks, "the procedure only involves a direct consideration of the ascribed belief's *content*" and the exercise of the same "normal abilities and dispositions for forming beliefs about the world" (1982, 225). (This, NB, despite the fact that "mastery of this procedure" does *not* exhaust a full understanding" of the self-ascription's *content* (1982, 226).)

As I read Evans, he takes it as given that ordinary self-ascriptions of beliefs (and other states) *are* instances of knowledge, and uses transparency to demystify their special status. His concern is to provide a non-Cartesian explanation of how conceptually articulate judgments concerning states of oneself can constitute *knowledge about* those states, even though they are not grounded in direct *consideration of* those states. But Evans' discussion leaves it open to different interpretations of his proposed "transparency procedure". In the remainder of this section, I canvass two recent construals of transparency-to-the-world – one *epistemicist* and one *metaphysicist* – both of which, in effect, deny the appearance of avowals'

groundlessness. Part of my aim in the following section will be to present my preferred, alternative reading.

2.1. *Byrne's epistemicism – transparency as inference*

In a series of articles, Alex Byrne has proposed that the best way to construe Evans is as pointing at a special but straightforward *epistemic procedure*: one

> *finds out* that one believes that it's raining by determining that it's raining: knowledge that one has this belief . . . *rests on* perceptual *evidence* about the weather, not on perceptual evidence of one's behavior or anything mental. That is, one *reasons* from evidence that it's raining, to the conclusion that one believes that it's raining. (2005, 93, first three emphases added)

Like other proponents of epistemicism, and in keeping with the presupposition mentioned earlier, Byrne tries to show that the very features that render transparent self-attributions knowledgeable will *also* explain the commonsense intuition that such self-beliefs are epistemically unusual and different from other attributions (including mental attributions to others and bodily self-ascriptions).

The main idea of Byrne's *inferentialist* view is as follows. In paradigmatic cases of belief self-attributions (as in Evans' "I believe there will be a third world war"), one arrives at the self-attribution via "an inference from world to mind: I infer that I believe that there will be a third world war from the single premiss that there will be one" (2011, 203). In other words, one reasons in accordance with the "doxastic schema":[12]

$$\text{BEL}: \quad \frac{p}{\text{I believe that p}}$$

A subject, S, reasons in accordance with the rule BEL when he or she believes that p *because* he or she recognizes that p obtains. The schema is intended to capture a "cognitive transition" the subject makes, or can make, which is not intended to be an explicit, conscious inference.[13] The "because" here – as in other epistemic rules Byrne considers – is the "because" of the epistemic *basing relation*: the subject's recognition that p serves as both the causal and the rational ground for her *belief that she believes* that p.[14] But reasoning in accordance with BEL is unusual in that the content of the mental state that the subject ends up believing she is in is identical to the content of the rule's world-directed antecedent.[15] In this way, BEL seems to capture the distinctive transparency of belief self-attributions.

Byrne acknowledges that the doxastic schema "is neither deductively valid nor inductively strong" (2011, 204); indeed, you might agree with a recent commentator who has protested that "only a madman could draw such an inference" (see Boyle 2011, 230). However, Byrne argues that the transition characterized by BEL *is* actually epistemically good. This is because BEL is not merely reliable; it is *self-verifying*: if S believes that she believes that p as a result of following BEL, then her belief that she believes that p *must* be true. (Byrne understands recognition in terms of "cognitive contact" with the relevant state of affairs, which is what is supposed to help explain why the rule BEL is *self-verifying*.[16]) This means that beliefs produced in accordance with BEL are *safe*, in the sense that they could not easily have been false (see Byrne 2005, 96–98, 2011, 206f.). So BEL seems to provide a promising model for epistemically good rules (or "schemata") that an

epistemicist can formulate to explain both the "peculiar" and the "privileged" character of transparent self-knowledge.[17]

Byrne offers his inferentialist account as an alternative to non-epistemicist accounts that use the transparency of belief to oppose "detectivist" views of self-knowledge.[18] In a recent response, Matthew Boyle has objected to Byrne's inferentialist alternative on several grounds. To begin with, he thinks Byrne's conception of inference as "merely a reliable process that deposits beliefs in my mind" is misguided. As a reflective person, "I can *reflect* on why I draw a certain conclusion, and when I do, I can see (what looks to me to be) a reason for it". But it is "hard to see how the premise of Byrne's doxastic schema could supply me with a *reason* to draw its conclusion" (Boyle 2011, 231, emphases added). More importantly, Boyle thinks that the very idea that, even in the normal case, we should need to rely on an *inference* from "sheer propositions" about the world in order to know our present states of mind "seems to get matters backwards" (2011, 234); it is as wrong-headed as the idea that we should need to *observe* their presence and character. Thus, he thinks that, in addition to relying on a dubious conception of inference, Byren's inferentialist presents a picture of basic self-knowledge that is as "profoundly *alienated*" as the "spectatorial" picture offered by the introspectionist.[19]

2.3. *Boyle's metaphysicism – making tacit self-knowledge explicit*

As an alternative to Byrne's inferentialist reading of Evans' transparency, Boyle offers a *reflectivist* reading, which construes

> doxastic transparency . . . as a matter . . . of shifting one's attention from the world with which one is engaged to one's engagement with it . . . an engagement of which one was already tacitly cognizant even when one's attention was "directed outward". (2011, 228)

The reflectivist maintains that "in the normal and basic case, *believing P and knowing oneself to believe P are not two cognitive states; they are two aspects of one cognitive state*" (Byrne 2011, emphasis added). Far from representing the culmination of an inferential step from a "sheer proposition" to a self-attribution, a subject's avowal "I believe P" represents "a coming to explicit acknowledgment of a condition of which one is already tacitly aware"; "to pass from believing P to judging I believe P, all I need to do is reflect – i.e. attend to and articulate what I already know" (Byrne 2011, 8–9).

Like proponents of the so-called constitutivist approaches to self-knowledge, Boyle thinks that no epistemicist account that appeals to an "information-gathering faculty" for producing beliefs about "an independent states of affairs" – whether inward *or* outward looking – could adequately capture the character and source of the special epistemic status of self-knowledge.[20] And like constitutivists, Boyle thinks that we can *only* avoid such appeal if we accept that the *facts* known when we know our minds are "ones whose holding is not independent of our being aware of their holding" (2010, 10). This *metaphysical* aspect of the problem of self-knowledge is highlighted when we look at the problem "from the first-person standpoint". What we need is "an account of what mental states *are* that explains how" it "can seem reasonable", from the standpoint of the avower, to make an avowal without any epistemic basis (2010, 16) and how an avower can "suppose that he is entitled to take the propositions he asserts to be true" (2010, 17).

Now, although I agree with Boyle's objections to epistemicism, I have some general misgivings about his Cartesian-style metaphysicist appeal to the essential knowability of

mental states, to which I will return in the final section. But even setting these aside, I find it very unclear how building self-belief into the very nature of mental states could help satisfy Boyle's demand for an explanation of what makes an avowal *reasonable* "from the standpoint of the person" who produces it. Briefly, if there really were for me (*pace* commonsense) a pressing question regarding my *epistemic entitlement* to my avowals, it is very unclear how "an account of what mental states *are*" by their nature could possibly help. Such an account (say it was handed down to me by God) might assure me that I cannot *be* in a mental state without knowing it (at least tacitly), but how would it allow me to think of my avowal as *reasonable* ("from my standpoint")? It seems to me that, from my standpoint, the appeal to the *metaphysical* nature of mental states would have no more probative *epistemic* value than an epistemicist appeal to the actual workings of my cognitive system and its various empirically reliable transitions. Thus, if we suppose (with Boyle) that the questions "What justifies my self-attribution? What makes it reasonable, *from my perspective*?" can indeed arise for me, even as I avow being in a certain mental state, then any epistemic assurance I might derive from a fact about the real nature of my mental state would seem no less alienated than that envisaged by an account like Byrne's. What is needed is an explanation of why the question of my avowals' reasonableness does not as much as arise for me in the first place, rather than an account of the metaphysical resources for answering it, should the question arise.

Be that as it may, anyone dissatisfied with both epistemicism and metaphysicism will want an alternative account that allows us to see how each of us is indeed in a special position to make knowledgeable pronouncements about our current states of mind, which, however, supposes *neither* that we normally rely on a special method of *detecting* the relevant states *nor* that our being in the states by itself implies self-knowledge.

3. Neo-expressivism: avowals' security and self-knowledge

Toward seeing our way clear to such an alternative, note that Boyle's metaphysicist rejection of Byrne's epistemicism is still motivated by an implicit acceptance of a presupposition he shares with the epistemicist. This is the idea that an account of the special character of self-knowledge must explain it in terms of *what renders mental self-beliefs justified*. The presupposition may seem natural, even if not inevitable, if we focus excessively on self-attributions of *belief* as the paradigm case of transparent self-knowledge. We can perhaps begin to free ourselves of the presupposition by keeping in mind, first, that transparency-to-the-world is not exclusively characteristic of avowals of doxastic states like beliefs (which are responsive to reasons and beholden to standards of rational justification). If asked (or when considering) whether one wants or prefers x, is annoyed at y, perceives z, plans to φ, and even remembers that q, one will typically answer by directly considering the intentional objects or contents of the relevant states.[21] So it is not *only* self-attributions of beliefs that are transparent in the relevant way; some avowals of non-doxastic states also enjoy transparency-to-the-world. But, perhaps more importantly, it is not plausible to suggest that *all* avowals partake in the transparency-to-the-world of beliefs (whether understood Byrne's *or* Boyle's way). For an especially telling example, consider solicited self-attributions of passing thoughts and unbidden desires. My authoritative response to the invitation "A penny for your thoughts" – for example, "Oh, I'm thinking about my grandmother" – could hardly be obtained through transparent consideration of relevant worldly affairs. And it is implausible to think of an unsolicited avowal such as "I'd love a cup of tea right now" as something *arrived at* through direct consideration of the

world (even though it shares content, as well as conditions of use, with a world-directed statement such as "A cup of tea would be nice right now"). Although in these cases the first-order state is itself directed at the world, one's *self-attribution* of it is not plausibly arrived at by *considering* the state's intentional objects. (For example, I do not tell *that* I am *thinking of* my grandmother right now by *considering my grandmother* – it is not even clear what that would amount to.)[22]

3.1. *The (Un)Myth of Jenny: avowals as expressive acts*

Focusing primarily on self-attributions of beliefs, both epistemicist and metaphysicist accounts suppose that the phenomenon of transparency (and the related phenomenon of Moorean anomalies) reveals that ordinary belief avowals must enjoy a special form of epistemic support. But once we recognize the scope and limitations of transparency, we may revise our understanding of its epistemic significance. Elsewhere, I have proposed that (what *Sellars* describes as) the "avowal role" of self-reports is the role of *expressing* the self-attributed states, rather than – or in addition to – one's second-order *belief* about the presence of the state.[23] On this proposal, in a sense, the transparency-to-the-world of self-attributions of beliefs falls out as but a special case or symptom of a broader phenomenon: the *transparency-to-the-subject's-state-of-mind* of avowals understood as expressive acts. I shall explain.

Let me begin with what – for purposes of juxtaposition with Sellars' Myth of Jones – we may call the (Un)Myth of Jenny. This is a rough-and-ready description of a familiar path taken by a young child, Jenny, from nonlinguistic behavior that naturally expresses her states of mind to articulate verbal expressions thereof. Seeing a fluffy toy – a teddy bear – little Jenny may stretch out her hand, eagerly reaching for the toy. The adult who recognizes what she wants might say "That's Teddy. You want Teddy, don't you?", in effect passing onto the child a new expressive vehicle for articulating aspects of the state she has expressed through nonverbal behavior. Subsequently, Jenny might give voice to her desire by exclaiming "Teddy!", with an accompanying gesture, then perhaps "Gimme Teddy!" in an eager tone of voice, and finally simply avowing "I want Teddy!". As performances, Jenny's eager reaching and her verbal utterances are of a piece; they are spontaneous, nonreflective or "unstudied", "excited" behaviors. But they employ different vehicles of expression. (The same holds for adult language users, who may express their amusement at a joke by laughing, as well as by saying "This is so funny!" or "I find this *hilarious*".)[24]

We can capture the relevant similarities and differences by calling upon a distinction also due to Sellars (1969).[25] The performances are similar in being *acts* in which an agent gives direct expression to a specific state of mind – expresses it *in the action sense* (a-expresses, for short), though in each case she is using a different *expressive vehicle*. A-expression is a three-place relation: an agent J a-expresses mental state M by using expressive means or vehicle E, where E can be bodily demeanor, facial expression, or gesture, whether natural, culturally acquired, or even idiosyncratic; it can *also* be a bit of verbal behavior. A-expression is to be distinguished from expression *in the semantic sense* – s-expression, for short – which is a relation that holds between contentful tokens, such as sentences, and their semantic contents. As we saw earlier, some expressive vehicles – laughter, for example – do not s-express anything. And one can use sentences that s-express different propositions to a-express one and the same state of mind.[26]

Borrowing an insight from traditional expressivism about avowals, my *neo-expressivist* account explains avowals' distinctive security in terms of their expressive character. The initial expressivist observation is that while anyone can say truly, and some can even tell

reliably, *that* I am feeling sad, only *I* am in a position to *express* my sad feeling itself – for example, by letting tears roll down my cheeks, or saying "This is so sad". Now, using the aforementioned distinctions, we can note that when *you* say "DB is feeling sad", you are employing a sentence that s-expresses the proposition that DB is feeling sad, and, if you are sincere, you are a-expressing your *belief* that DB is feeling sad. My tears, on the other hand, s-express nothing (though *I* may be a-expressing my sadness by letting them roll down my cheeks). Now consider the sentence "This is so sad". It s-expresses a proposition that describes something as sad; but in uttering it, one would typically be a-expressing her sadness. Similarly, according to the *neo-expressivist* account I advocate, when avowing "I feel sad", one performs a distinct type of act which serves directly to a-express a present state of mind – *the very state that is self-scribed by the proposition that the sentence used s-expresses. As expressive acts*, avowals – like non-self-ascriptive expressions (including natural expressions) and unlike evidential reports (whether third- or first-person) – are protected from correction and demands for reasons or justification. However, insofar as avowals use as expressive vehicles truth-evaluable sentences that s-express self-ascriptive propositions, they are importantly different from other kinds of expressions of the relevant states.[27]

Now, on the Sellarsian Theory-theory, we saw, an avower is someone who has learned to respond to her own occurrent mental episodes by spontaneously and non-inferentially producing an avowal "out loud" or silently. I am proposing, instead, that an avower is someone who has acquired articulate, self-ascriptive vehicles by way of replacing, or supplementing, other expressive vehicles and for whom giving articulate voice to her states of mind has become "second nature". But, if we are to understand the "avowal role" on the model of expressive acts, we need some understanding of what holds together expressive acts that use diverse expressive vehicles.

In recent work, I have explored the idea that acts of expressing one's state of mind involve using vehicles that are in some sense *designed to show* that state (see, e.g. Bar-On 2004, Ch. 7, 2013a, 2013b). An animal baring its teeth in anger, or suddenly shifting its gaze, a child smiling in pleasure, a person raising an eyebrow, or even blurting out a curse, all engage in spontaneous behavior – whether unlearned or acquired – that "springs directly" from a state of mind they are in, and is neither preceded by deliberation nor even mediated by communicative intentions. And the behavior directly reveals to suitably endowed recipients the state of mind it expresses. The idea is intuitively plausible when we think of natural expressions. The capacity to spontaneously engage in and immediately recognize behavior that is naturally designed to show states of mind is a capacity with deep phylogenetic roots that connects us with nonhuman creatures as well as with our younger and nonreflective selves.[28] But this idea, I think, has application that goes beyond natural expressions, which, as expressive vehicles, are unlearned or innate. Linguistic exposure, habituation, enculturation, and other social experiences enable expressers to integrate into their expressive repertoires a wide variety of acquired ("nonnatural") expressive vehicles with which to give voice to their states of mind. Thus, in creatures like us, some of the communicative roles played by the more "visceral" showing and perceiving afforded by animals' growls, bared teeth, grimaces, and so on is taken up by spontaneous, competent use and immediate uptake of linguistic vehicles. (Swear words are one good example, but by no means the only one.) So nothing in the transition to verbally articulated expression requires, as far as I can see, retreating to the idea that verbal expressions enable knowledge of expressed states *only* in virtue of speakers intending to be providing evidence of – and their witnesses making inferences about – the presence of the expressed state.[29] Especially when it comes to avowals, it might be argued that the burden of showing and

immediately recognizing expressed states and their intentional contents is shouldered by the *semantic* features of the self-ascriptive expressive vehicle; for avowals *wear the conditions they are supposed to express on their linguistic sleeve*, as it were. An avowal such as "I wish we'd get some rain today" *explicitly names a kind of condition* (a hope) *and articulates its content* (*that it rain today*), as well as *ascribing it to a certain individual*; it reveals the *kind* of state the avower expresses (as well as its intentional content, when it has one) through what the sentence expresses in the *semantic* sense. (Contrast: "Rain would be great!" Or: "Oh for some rain!" which may equally serve to a-express a subject's wish for rain, but without *naming* the state.)[30]

3.2. *Expression and transparency*

Insofar as all acts of avowing can be said to express – and thus to show – the ascribed states (in virtue of the self-ascriptive expressive vehicles they use), avowals can be said to enjoy a certain transparency – what I elsewhere describe as *transparency-to-the-subject's-state.*[31] This is a different notion from Evans' notion of transparency-to-the-world discussed earlier, which pertains (at most) to avowals that explicitly specify some worldly matter or object "outside" the subject. For, in the relevant sense, *all* avowals are transparent-to-the-subject's-state, regardless of whether what is avowed is a phenomenal or intentional state, and whether or not the avowal concerns a state that is itself rationally evaluable. Moreover, on the neo-expressivist account, transparency-to-the-world falls out as a consequence of the expressive character of all avowals. If asked (or when considering) whether you believe p, you will normally directly attend to whether p is the case. We can think of this as a way of putting yourself in a position to give direct voice to your (first-order) belief, which is what the neo-expressivist account says you do when avowing. This is so whether you *already* have the relevant belief or are now forming it. What is important is that in neither case you need to *discover* what you believe. Instead, you simply give it voice, having considered (or reconsidered) whether things are as the proposition says. You pronounce on the truth of the proposition, though you are using a self-ascriptive expressive vehicle. But even unprompted, spontaneous pronouncements – such as "I'd like some tea", or "I'm wondering what time it is" – that are not preceded by "direct consideration of the world" can still be seen to partake in the expressive transparency of avowals.[32]

On the commonsense view, avowals are strongly presumed to be true, while at the same time being protected from ordinary epistemic challenges, demands for reasons, justification, etc. The neo-expressivist account explains this by portraying avowals as acts of *speaking one's mind* self-ascriptively, *in lieu of* giving either nonlinguistic or else non-self-ascriptive expression to one's state of mind. When avowing, one acts so as to give voice, out loud or silently, to the very same state of mind that is named by the expressive vehicle that the avower uses in her expressive act. Thus, avowals, like expressive acts more generally, give direct voice to the avower's states of mind and allow others to *see through* to those states, though they do so using self-ascriptive vehicles.[33] And this explains avowals' remarkable epistemic features (namely, their epistemic ungroundedness as well as security).

3.3. *Avowals' security and self-knowledge*

Now, in keeping with common sense, and in contrast with epistemicist views, neo-expressivism tries to accommodate the idea that avowals are groundless. But it does not endorse the metaphysicist claim that, *in general*, being in a mental state *ipso facto* involves having knowledge of it. Rather, it maintains that it is misconceived to think of avowing subjects as

standing in any need of justification (inferential or otherwise) or positive reasons for taking themselves to be in the state that their avowal self-ascribes. Yet *pace* deflationism about self-knowledge, this does not entail that avowals cannot represent articles of (privileged) self-*knowledge*.

So, what does qualify avowals as articles of (privileged) knowledge? (Our earlier question (ii).) My preferred response to this question takes its initial inspiration from Evans. Evans identifies a form of epistemic resilience that he (and Shoemaker) describes as *immunity to error through misidentification*.[34] Considering certain bodily self-attributions – for example, proprioceptive and kinesthetic self-reports (as well as perceptual self-ascriptions such as "I see a canary") – he notes that they are "identification-free", in the sense that they do not rest on a recognitional identification of the subject of the utterance or thought. In normal circumstances, if I say (or think): "My legs are crossed", my self-ascription does not rest on my separate recognitional identification of someone as being me, and whose legs I also take to be crossed. Immunity to error à la Evans is essentially a *negative* notion: it does not entail freedom from all error – only freedom from a *certain kind* of error and protection from certain kinds of epistemic challenge. And, though it affords the relevant judgments a certain epistemic security, that security is not a matter of some specific positive justification. I have proposed, similarly, that avowals enjoy – in addition to immunity to error through misidentification – immunity to error through misascription, which immunity, in turn, can be explained by appeal to their expressive character. (Very briefly, the idea is that, when avowing, it is not only that I have no independent reason for thinking that it is *me* who is in the relevant state (which is what renders the avowal immune to error through *misidentification*.). It is also the case that whatever reason (or, better, warrant) I have for ascribing a particular mental state to myself, it does *not* derive from a separate reason I have for believing that *someone* is in the relevant mental state, *or* that I am in *some* state or other, or that the state concerns *something* or other. So my avowal is *also* immune to error through *misascription*.[35])

However, what's more important for our present purposes is that Evans separates the epistemic resilience afforded by immunity to error from the positive features that render judgments that enjoy such immunity *knowledgeable*. Thus, if I judge in the normal way that I am sitting down, what renders my judgment that it is *me* who is sitting down immune to error through misidentification is the *absence* of reliance on some means of recognizing myself as the one who is sitting down. But this absence is *not* what *explains* my *knowledge* of the relevant fact. On Evans' view, the explanation of my knowledge that it is *me* who is sitting down comes from the fact that we, human beings, possess "a general capacity to perceive our own bodies" (which includes "our proprioceptive sense, our sense of balance, of heat and cold, and pressure"), as well as a capacity for determining our own "position, orientation, and relation to other objects in the world … upon the basis of our perceptions of the world" (1982, 220, 222). It is *the exercise of these capacities* for gaining information about some of our states that allows the relevant self-judgments to represent a certain kind of bodily self-knowledge *despite* the fact that they do not involve self-recognition (and are thus immune to error through misidentification).[36]

Along similar lines, I have proposed that the *positive* story about what avowals have "going *for* them", epistemically speaking, should be told separately from the story about their distinctive security. Indeed, I believe there are a number of positive epistemic stories we can tell, consistently with the neo-expressivist explanation of avowals' security.[37] Given this explanation, though, and following Evans' lead, it would be natural to propose

that it is *the exercise of our capacity to give voice to our present states of mind* – to speak our minds – that allows avowals to represent mental self-knowledge.[38]

4. Expression, self-knowledge, and folk psychology

My proposal, then, is that the epistemic achievement involved in ordinary self-knowledge is due to the fact that it represents the exercise of a unique capacity we each have to express our present states of mind, using truth-evaluable, self-ascriptive expressive vehicles. I think this proposal ties nicely with our *psychological* understanding of ourselves as creatures who are capable of showing how we feel, what we want, what we are attending to or interested in, etc., to suitably placed others. The capacity for expressing states of mind is the prerogative of all genuinely social species, and may even be coeval with genuine mindedness. Regarding ordinary, basic self-knowledge as growing out of expressive capacities comports with an important feature of the commonsense view of the basic relation between the first-person and others, which is arguably rather different from the traditional Cartesian picture. As subjects of experience, we are *in* various states of mind, and can express them through our behavior. A subject in the grip of anger, fear, anxiety, or even just focused attention will typically show it through some facial expression, gesture, bodily demeanor, eye movement, etc., thereby immediately betraying his or her state of mind to suitably endowed observers. On the neo-expressivist view – as on the folk-psychological picture – we do not merely perceive others' *behavior* and *infer* to the presence and character of their states of mind as the best explanation of the behavior we perceive; rather, we immediately recognize their present states of mind *through* their behavior.

Crucially, however, unlike nonverbal creatures, we – normal adult human beings – can give articulate voice to our mental states – we can *speak* our minds. As we saw, this is an acquired ability; our Jenny has learned to articulate aspects of states of mind by being handed down linguistic vehicles of expression. Still, the neo-expressivist insists that the acquired ability to speak one's mind is an ability that emerges out of and builds upon other ways of showing states of mind, and reflects various psychological continuities between us – normal, adult human beings – and nonlinguistic creatures, as well as our pre-linguistic selves and differently abled humans.

I see this as an advantage of neo-expressivism over views that postulate a constitutive link between having, for example, a belief and knowing that one has it (even if tacitly). As common sense would have it, being in a mental state is one thing, *knowing* that one is in the state is another. Our folk psychology licenses attributing various (first-order) mental states to creatures (nonhuman, very young humans, or very challenged humans) who do not possess the requisite reflective capacity or concepts necessary for having higher-order *beliefs about* states of mind. They may lack *self-knowledge,* but for all that, it is much less obvious that they lack all first-order states of mind. Yet, as we saw earlier, Boyle stipulates that privileged self-knowledge is a matter of *reflective attention* to states that *are by their nature already tacitly known*. Since availability to reflective attention requires possession of the relevant concepts, the reflectivist has to insist that *any* attributions of beliefs (and other mental states) to nonreflective nonhuman creatures must be understood as made "in a different register". For the reflectivist – as for constitutivists more generally – "brutes" (as well as prereflective humans and adults whose reflective capacities are diminished or impaired) can at best be said to have beliefs in an attenuated sense.[39] But even as regards fully formed, normal, reflective humans, it seems that the reflectivist would need to relegate allegedly unconscious emotions, wishes, thoughts, feelings, and beliefs that are unnoticed or not actively reflectively attended to, and so on, to a second class, "lower

case" mental status.[40] This commits reflectivism to a rather extreme bifurcation – what I have elsewhere dubbed *Mind-mind dualism*: the separation of even human psychology into two distinct realms: the "merely psychological" realm, the denizens of which are at best dispositions, or passive-responsive occurrences that only have powers to move us *causally*, on the one hand, and the genuinely Mental realm, which is thoroughly normatively governed and essentially reflective, on the other.[41]

I agree with Boyle that, in creatures that *are* capable of self-attributing mental states, the relationship between first-order mental states and their avowals is more intimate than would be suggested by epistemicist views (though how to characterize the relationship is a delicate matter that goes beyond my scope here). But I think that building reflective accessibility – which is a rather demanding achievement conceptually as well as epistemically – into the very nature of mental states strays equally from our folk-psychological conception of our relation to our states of mind as reflected in the role that avowals ordinarily play in our lives. Adhering to this conception, the neo-expressivist view maintains that avowals are importantly continuous with other ways we have of giving direct voice to our states of mind and of allowing others to see through to them.

As I have stressed throughout, though, avowals differ from other expressions in that they involve *speaking* our minds, using self-ascriptive expressive vehicles. Importantly, learning to speak one's mind is not just a *linguistic* achievement; it is also a *psychological* achievement. This can be appreciated when things go wrong, and a subject is *not* able competently to avow her feelings, thoughts, beliefs, preferences, wants, etc., or else misexpresses his or her states of mind. Expressive behaviors in general – and not only verbal expressions – are subject to a range of expressive failures. A subject may not only fail to express (whether deliberately or not) a state of mind he or she is in; but he or she may also say, for example, "I like this painting", sincerely, and without sarcasm or dissimulation, despite in fact *not* liking the painting. Subjects' avowals can be false in cases of self-deception, wishful thinking, implicit biases, and a host of other psychological irregularities. Citing such irregularities, as studied by psychologists, many have argued that our commonsense conception of basic self-knowledge as privileged is seriously misguided. Various experiments are said to establish the "opacity of mind" (Carruthers 2011), the "unreliability of naïve introspection" (Eric Schwitzgebel), and the fact that, in reality, we are "strangers to ourselves" (Timothy Wilson).[42] Yet, as I see it, the neo-expressivist approach, with its "divide and conquer" strategy, is equipped to handle such challenges – and better than the epistemicist and metaphysicist views discussed earlier.

To see why, note that there is an important difference that sets the neo-expressivist account apart from both the inferentialist and the reflectivist views canvased earlier. On both these views, avowals' security derives directly from whatever it is that confers on them positive epistemic status – what *justifies* or gives us reasons for the relevant self-beliefs. This means that any systematic evidence showing that subjects can self-ascribe states they are not in, or can be in states they fail to self-ascribe, would directly compromise avowals' security as understood by these views. But such evidence, I submit, does *not* directly threaten avowals' security as understood in terms of avowals' expressive role.

Take, for example, the kind of experiments cited by Carruthers, in which experimenters solicit subjects' opinions on certain matters, where it is clear that the opinions they provide have been influenced by extraneous factors, as evidenced by the fact that they are inconsistent with opinions they had given previously, *and* where there is no reason to suppose that they had changed their minds. In some of the experiments, subjects are induced to write an essay arguing for a conclusion – for example, that a tuition increase would be acceptable –

that is contrary to what they believe (as evidenced not only by prior statements, but also by, say, various actions they take). Those subjects who are led to believe they freely chose to write the essay then go on to self-attribute the opinion they had defended in the essay. Carruthers argues that "the best explanation" of the patterns of results in these experiments is that "subjects' mindreading systems automatically appraise them as having freely chosen to do something bad resulting in negative affect" and then when asked for their opinion, they "select" the response that "provides an appraisal of their actions as being significantly *less* bad". However, by Carruthers' own characterization, these experiments do *not* provide evidence of the subjects' *self*-attributions – their avowals – pulling apart from non-self-attributive expressions of their first-order state. A subject in the experiment who avows "I believe tuition increase would be acceptable" would *also* pronounce (or assent to) "Tuition increase would be acceptable". But, I submit, so long as the subjects' avowals line up with other non-self-ascriptive expressions of their propositional attitudes, avowals' security as the neo-expressivist understands it remains intact. Note that I am here going along with Carruthers' assumption that the subjects have *not* changed their beliefs as a result of the experiment, so that the subject's ("real") beliefs are indeed out of synch with her avowals. This is also the common understanding of "implicit bias" experiments – where subjects' biased behavior and actions are alleged to give the lie to their avowals of impartiality. My point is that insofar as the beliefs are *also* out of synch with his or her non-self-ascriptive pronouncements, the incongruity in the subject's "system" is one that does not have to do, specifically, with *self-attribution*. So the experiments do not show compromise in the security of avowals *as understood on the neo-expressivist account*.[43]

As in humdrum cases of so-called self-deception, when subjects in psychological experiments self-attribute beliefs, opinions, preferences, etc., that they do not have, clearly something is awry, psychologically speaking. But I submit that what goes wrong does not tell *specifically* against the reliability of *self*-attributions or the authority of those issuing them. The subjects' avowals are false, but their falsity is not due to an *epistemic mistake* on the subjects' part (namely, failing correctly to recognize or reflectively attend to a state they are in). Instead, the falsity represents an *expressive failure*, where the failure has identifiable *psychological* causes. (That the failure is psychological, rather than a failure of knowing one's mind is, again, evidenced by the fact that, under the circumstances, the subjects' non-self-ascriptive expressions will match their avowals. There is, of course, a good question as to the causes of the mismatch between the subject's *non*-self-ascriptive expressions and his or her states of mind. But the point is that these should be seen as *psychological* causes that do not bear exclusively on the subject's ability to *self-attribute* those states.[44])

To reiterate, a subject who, due to some extraneous influence, avows "I think a tuition increase would be ok" would be equally disposed to say confidently: "A tuition increase would be fine" (just as a subject on a dentist chair who falsely avows a toothache would be equally disposed to wince). So there is, to borrow a phrase from David Rosenthal,[45] a general "performance equivalence" between self-ascriptive avowals and "first-order" expressions of beliefs (and other states). It is this performance equivalence that, in the case of (at least some) propositional attitudes, manifests itself as transparency-to-the-world, and which neo-expressivism explains by appeal to the expressive character of all avowals. Given the neo-expressivist understanding, it would make no more sense to think of the mistaken avowal as a product of mistaken self-inference, or of self-reflection gone awry, than it would to think of a non-self-ascriptive expression of belief as such a product. Likewise, unless we are willing to accept the idea (absurd, I would submit) that

all of our expressive behaviors are products of self-interpretation, we should not suppose that the *self*-attributions are such products.

The performance equivalence just mentioned does reflect an intimate connection between first-order states and self-attributions of them. But, while I think it is a mistake to dismiss the connection as a merely superficial feature of the *pragmatics* of mentalistic discourse, I do not think it lends support to either the epistemicist view or to the view that mentality is *essentially* tied to self-knowledge. Rather, it is a consequence of the more mundane – which is not to say trivial – fact that our so-called first-person authority emerges out of a psychological capacity we share with other minded creatures to express our present states of mind, coupled with the special capacity we have as linguistic creatures to use articulate, self-ascriptive vehicles to do so.

Acknowledgements

I wish to thank Fleur Jongepier, Kate Nolfi, Tory McGear, and Carol Voeller, as well as the audience at the conference on Self-Knowledge and Folk Psychology, Radboud University, Nijmegen, Netherlands, 27–28 June 2014 for helpful discussions and comments.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes

1. I discuss avowals' epistemic security and offer an account of it in*, inter alia*, Bar-On (2000, 2004, 2009a), and Bar-On and Long (2001).
2. See, e.g. Sellars (1956, XV.56, XV.59). For an account of introspective privileged access very much along these lines – in terms of reliable second-order non-inferential (and "silent") responses to one's own first-order mental states – see Rosenthal (2005). For helpful exposition, see O'Shea (2007, 92ff.).
3. In my final section, I will briefly present an alternative to Sellars' myth of Jones, one inspired by our real, non-mythical lives.
4. I here follow the rough division in Gertler (2011, Ch. 1) of contemporary views into empiricist vs. rationalist.
5. In keeping with this way of thinking, it is often assumed that
   (a) our folk psychological conception of avowals' distinctive security reflects our taking them to be prime instances of knowledge, knowledge that is, moreover, privileged and
   (b) (therefore?) any explanation of avowals' distinctive security must appeal to whatever features render them instances of knowledge.
6. See, e.g. Bar-On (2004, Ch. 1, 2010, 2012; Bar-On and Long 2003). For more recent discussion, see Bar-On and Nolfi (forthcoming).
7. Parts of this section overlap Bar-On and Nolfi (forthcoming). I wish to thank Nolfi for permission to use these materials and for helpful discussions.
8. I am here focusing on *epistemic*, as opposed to psychological groundlessness (see e.g. Cassam 2014). For my own take on the distinction, see Bar-On (2004, Ch. 6 and *passim*).
9. Inner sense, recall, pins the security of avowals and first-person authority to the high degree of reliability – even if not Cartesian infallibility – of our faculty of inner perception; and acquaintance theories pin it to the existence of an unmediated relation of direct acquaintance we have to our own current states of mind. For a survey, see Gertler (2011, Ch. 4, 5).
10. Compare Boyle (2011, 226).
11. For the reading of transparency that follows, see Bar-On (2000) and Bar-On (2004, Ch. 4); see also Bar-On (2009a, 2010, 2012).
    Following Evans, I prefer to speak of transparency-to-the-world as a feature of (some) *self-ascriptions of beliefs* and other states of mind (that is, second-order judgments on our first-

order mental states), rather than speaking – as do some recent authors – of our first-order *beliefs* (and other states) *themselves* as enjoying transparency. For discussion, see Bar-On (2009b). (This will become relevant later on.)

12. Byrne (2005) talks of the "epistemic rule" BEL: If p, believe that you believe that p – which he formulates to fit epistemic rules of the general form R: If conditions C obtain, believe that p. (He considers, for example, DOORBELL: If the doorbell rings, believe that there is someone at the door. Or NEWS: If the Weekly News reports that p, believe that p (see Byrne 2005, 94)).
13. Given the way Byrne conceives of reasoning in accordance with BEL, it seems legitimate to question whether what is at work deserves to be described as genuine *reason*ing, with the premises representing the *subject's reasons for* believing the conclusion, or instead just a (sub)cognitive "movement of the mind" from one cognitive state to another (see Boyle 2011, 7–9).
14. Also crucial in Byrne's formulation here is that recognition entails knowledge. Thus, the fact that a subject recognizes that the antecedent obtains entails that the subject knows (and so also that the subject believes) that it obtains.
15. This, however, may be a peculiarity of the case of belief; see the following.
16. However, there are cases in which one recognizes p, but also affirms that one does not believe p (or believes not-p) based on, for example, a therapist's analysis, or interpretive self-analysis (see below). So cognitive contact with p does not seem sufficient for transparency; for the attribution to be transparent, one must *attribute* the belief in a certain way.
17. Byrne (2011) proposes to extend the account to, for example, the case of self-knowledge of intention, by appealing to the epistemic rule:

INT I will ø
– – – – – – –
I intend to ø

18. See Byrne (2011) for relevant references and objections.
19. Bar-On (2004, Ch. 4) raises a closely related objection to the Epistemic Approach.
In the case of a rule like INT (see endnote 17), the idea Boyle regards as misguided is that "I might infer propositions about my present intentions from blank future propositions about myself from a blank future proposition about myself, as if I must conclude my own commitment to ø from an unaccountable inkling about what I will in fact do" (2011, 234).
Boyle thinks that Byrne's inferential approach to doxastic transparency faces a certain dilemma: it must either represent the subject as drawing a mad inference, or else must admit that her real basis for judging herself to believe P is not the sheer fact that P, but her tacit knowledge that she believes P (which would effectively mean giving up on Byrne's project) (see Boyle (2011, 231f.).
20. For some discussion of constitutivism (and relevant references), see Gertler (2011, Ch. 6), Coliva (2012), and Bar-On (2009a). Boyle's metaphysicism has several important points of contact with constitutivist views; however, Boyle distances himself from constitutivism (cf. 2011).
21. Although as I argue in Bar-On (2004, 118f.), direct consideration of the relevant worldly items is not an especially reliable method for determining one's desires, preferences, and other states. For example, considering the bulldog in front of me, I may judge that it is not to be feared, yet I may feel very scared of it.
22. Thanks to Tory McGear for prompting this clarification.
Similar remarks apply to Moore's paradox, which is often mentioned in connection with transparency. Notably, transparent consideration of outward phenomena can issue in self-ascriptions that are ripe for being caught in Moorean absurdities. But there are many examples of Moore-style anomalies that do not involve beliefs (consider, e.g. "I'm finding this meeting really exciting, but it's very boring"; "Tea please! But I don't want any tea"; and even "Brrr! It feels hot in here"; or "[Agonized expression] I feel so happy"). Moreover, even focusing on belief, Moorean conjunctions *can* be rationally produced or entertained. (For discussion, see Bar-On 2009b.)
23. I develop this view in (2004, esp. Ch. 6–8) and elsewhere. For a defense of a somewhat similar expressivist view on the basis of considerations from autism, see McGear (2004).
24. For my most recent summary of distinctions I am using here and in the following, see Bar-On (2015).

25. However, to be clear, I am not endorsing Sellars' account of *self-knowledge*, but instead seek to articulate an alternative to it.
26. I here set aside, for the most part, what Sellars (1969) calls (misleadingly, I think) "expression in the *causal* sense" – for example, nonvoluntary, uncontrolled facial expressions or gestures that reveal one's state of mind. This is because the expressive behaviors relevant to my concerns here – avowals – are *not* nonvoluntary or reflexive bodily happenings, but rather things that are *done by an individual* (as opposed to a subsystem, or module, within the individual), over which the individual exercises a certain kind of central, executive control (see Bar-On 2004, 216f., 249ff., 289, 315).
27. Thus, like traditional avowal expressivism, neo-expressivism does not regard avowals' distinctive security as inherited from the security of this or that epistemic basis on which they are made or from the very nature of mental states. But, unlike the traditional view, neo-expressivism emphasizes important dissimilarities between avowals and inarticulate grunts, grimaces, or cries, and even many verbal expressions. Unlike the latter, avowals exhibit various *semantic continuities* with other attributions of states (both to oneself and to others). (For a full development of the account of avowals' security – the epistemic asymmetries, as well as the presumption of truth governing them – see Bar-On 2004, Ch. 6–8; in keeping with my rejection of the presupposition mentioned earlier, the account of what could render avowals articles of *knowledge* is provided separately, in Ch. 9.)
28. As an aside: of course, on a given occasion, an animal can bare its teeth without being angry. This can happen for any number of reasons, and not necessarily because the animal is trying to deceive. In such cases, of course, the teeth-baring does *not* allow us to perceive the animal's being angry (by perceiving its baring its teeth). Showing and perceiving are both *factive.* But that does not mean that we cannot apply the idea of perception in such cases. (For discussion, see Bar-On 2004, 240ff., 310ff., and 410ff.)
29. For relevant discussion, see Bar-On (2015).
30. Unlike in the case of natural expressions, the relevant information is not revealed through perception-enabling features of the expressive behavior. It is made available to us through the linguistic vehicle used in the act of avowing.
31. See Bar-On (2004, 310ff.). The relevant transparency is enjoyed by avowals *qua* expressive acts, and is shared by all other such acts – it is due to the fact that expressive acts *show* the expressed state. I distinguish, in this connection, between, for example, expressing *pain* and expressing *my* pain, thereby making room for the possibility of *expressive failures* (see esp. Ch. 8).
32. The commissive aspect of transparent self-attributions of belief (and other reflective states) emphasized by reflectivist views can also be seen as a consequence of avowals' expressive character. Insofar as avowing a belief serves to give direct expression to the self-ascribed belief itself (rather than – or perhaps in addition to – one's second-order belief), in avowing (as opposed to merely reporting one's belief) one incurs commitment to the truth of one's first-order belief (see Bar-On 2004, 318–310, 2009a).
33. It may be thought that this shows transparency-to-the-world *not* to be a special case of transparency-to-the-subject's-state, since the former is supposed to be a "first-person" feature of certain self-attributions, whereas the latter is, in a sense, "third-person" (thanks to Tory McGear). However, this ignores the first-person epistemic consequences of taking avowals to be a species of expressive acts (for my take on the first- and third- person aspects of the neo-expressivist account, see my Bar-On (2004, esp. Ch. 7, 10).
34. See Evans (1982, esp. Ch. 7, Sec. 2), Shoemaker (1968), and Wright (1998, 18–20). For the relevance of immunity to error to the present topic, see esp. Bar-On (2004, Ch. 4, 6, 2012).
35. For discussion, see Bar-On (2004, Ch. 6, 2007, 2012).
36. Evans' construal of transparent self-ascriptions can be regarded as following the same pattern. For he portrays them as judgments that are *about* one's intentional states, but ones made *not* on the basis of introspective identification of a (self-ascribed) state inside us (as he says: "[With this] method of self-ascription . . . we . . . have no need for the idea of an inward glance" 1982, 225). Nonetheless, he takes a person's transparent self-ascription to represent "*knowledge of* one of his mental states: even the most determined sceptic cannot find here a gap in which to insert his knife", Evans (1982, 225). But it should be kept in mind that, unlike proprioceptive and kinesthetic self-ascriptions, transparent intentional self-ascriptions, as Evans himself notes, do not utilize any "special faculty of inner sense or internal self-scanning" (1982, 230, fn. 42). I

distinguish two questions in this connection: the question how one can make a *genuine* ascription of an intentional state, and the question why one is assured to make a *correct* ascription. Evans does not explicitly separate the two questions. The procedures he offers for explaining self-knowledge of intentional states seem to play both the role of explaining how one can think about one's intentional state at all (even though one is not directing attention at the state) and how one can obtain reliable self-knowledge regarding such states.

37. For discussion of various substantive responses to question (ii), see Bar-On (2004, Ch. 9).
38. For relevant discussion, see Bar-On and Nolfi (forthcoming).
39. Boyle wishes to dissociate himself from the label "constitutivism", but for reasons that do not bear on the objections here (see 2011, 228–229, fn. 5). In 2010, Boyle speaks of brute beliefs as constituting a *different species* of belief (though they may belong to the same genus as our reflective beliefs). Interestingly, in 2011, he says that he does "not wish to deny that creatures incapable of reflection can have belief" and takes "no position on whether their believing involves tacit knowledge of believing" (228, fn. 5). But, given that (even tacit) knowledge presumably requires belief, and given the plausible assumption that at least second-order belief requires the *concept of belief* (and of other mental states), it is difficult to see how he can avoid taking a (negative) position on the question at hand.
40. See Bar-On (2009a). Some constitutivists explicitly restrict their thesis to what they term *rational* beliefs and intentions, "judgment-sensitive" wishes, desires, and preferences, intentional states understood as *commitments*, and so on (see, e.g. Moran 2001; Bilgrami 2006). Alternatively, one can insist on judgment sensitivity and reason responsiveness as necessary conditions on a state *being* one of *belief*, *intention*, *desire*, etc., and then make Boyle's move (see previous footnote) of acknowledging a broad genus of which belief, intention, desire, etc., *proper* and their non-reflective analogues are both species. Either way, one must embrace what I go on to call *Mind-mind* dualism.
    In the case of reflective creatures, Boyle might insist that "being already tacitly known" does not entail "being reflectively attended to", thereby making room for the range of states under consideration. But this would seem to me to build much more into reflective attention than Boyle should allow, given his objections to epistemicism. Space limitations prevent me from elaborating on this worry here. (For relevant discussion, see Bar-On 2004, Ch. 9, 2009a); see also my reply to Boyle in 2010 (which I develop further in a paper in progress).
41. Of course, none of this is to deny that reflective beings like ourselves have, in addition to the types of states of mind we share with brutes, also types of states of mind that we do not share with them. And part of the difference may well have to do with the exercise of our reflective, self-transformative capacities.
42. Unlike Byrne and Boyle (and others) who try to accommodate and explain the asymmetry between attributions of propositional attitudes to oneself and to others, Carruthers (2011) endorses a "full symmetry" view, according to which we know all of our own propositional attitudes in exactly the same ("interpretive") way that we know of others', and we are equally prone to mistake in our mental self-attributions. Carruthers claims that psychological experiments show that people can attribute "thoughts to themselves that we have every reason to believe they never entertained, and making errors in self-attribution that directly parallel the errors that we make in attributing thoughts to other people". But he also thinks that the "studies show that … people are using the same mindreading faculty that they employ when attributing thoughts to other people, relying on sensory forms of evidence that stands in need of highly fallible interpretation", suggesting that "there is just a single mental faculty (the 'mindreading' faculty) that is responsible for all our knowledge of propositional attitudes, whether those thoughts are our own or other people's". (Others, like Schwitzgebel, have extended the scope of the claim of our radical fallibility to sensations.)
43. Thus, I submit, the experiments cannot help establish the claim that there's "full symmetry" between mental self-attributions and attributions to others; nor do they support his idea that avowals of propositional attitudes are the result of mindreading self-interpretation (as Carruthers maintains; see previous endnote). Similar remarks apply to a number of cases discussed by Nisbett and Wilson (1977) including ones showing "implicit bias") as well as to Schwitzgebel's "armchair" experiments concerning visual experiences. (The latter experiments are characterized in terms of asking subjects when they *see* an object in their visual field. But presumably the subjects would give the same answers if asked: "Tell me when *the object* appears".)
    If (as suggested to me by McGear) we insist that the subjects in the experiments do have the

beliefs they avow (and that are expressed by their non-self-ascriptive pronouncements), then, of course, no threat to avowals' security would arise. At least in some cases such insistence would require severing the connection of belief with action and insisting on too strong a connection between beliefs *qua* psychological states and verbal expressions of beliefs. However, this issue merits further discussion, which I cannot undertake here. (In particular, it is important to distinguish behavior and actions that are *caused by* one's beliefs and behavior (nonverbal or verbal) through which one *expresses* one's belief. See my 2004, Ch. 6–7 for relevant discussion.)

44. Thanks to Fleur Jongepier for prompting this clarification.
45. See Boyle (2010) and Bar-On's reply (2010).

Notes on contributor

Dorit Bar-On is Professor of Philosophy at the University of Connecticut (previously at UNC-Chapel Hill) and director of the research group Expression, Communication, and Origins of Meaning. She is the author of *Speaking My Mind: Expression and Self-Knowledge* (OUP 2004). She is currently working on a book manuscript tentatively titled *Expression, Action, and Meaning*, as well as a volume in the Wiley *Great Debates in Philosophy* series (with Crispin Wright).

References

Bar-On, Dorit. 2000. "Speaking My Mind." *Philosophical Topics* 28: 1–34.
Bar-On, Dorit. 2004. *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Calrendon Press.
Bar-On, Dorit. 2009a. "First-Person Authority: Dualism, Constitutivism, and Neo-Expressivism." *Erkenntnis* 71: 53–71.
Bar-On, Dorit. 2009b. "Transparency, Epistemic Impartiality, and Personhood: A Commentary on Simone Evnine's Epistemic Dimensions of Personhood." *Philosophical Books* 50 (1): 1–14.
Bar-On, Dorit. 2010. "Précis of *Speaking My Mind: Expression and Self-Knowledge* and Reply." *Acta Analytica* 25 (1–7): 47–64.
Bar-On, Dorit. 2012. "Externalism and Skepticism: Recognition, Expression, and Self-Knowledge." In *The Self & Self-Knowledge*, edited by Annalisa Coliva, 189–211. Oxford: Oxford University Press.
Bar-On, Dorit. 2013a. "Origins of Meaning: Must We 'Go Gricean'?" *Mind & Language* 28: 342–375.
Bar-On, Dorit. 2013b. "Expressive Communication and Continuity Skepticism." *Journal of Philosophy* CX (6): 293–330.
Bar-On, Dorit. 2015. "Expression: Acts, Products, and Meaning." In *Meaning Without Representation: Essays on Truth, Expression, Normativity, and Naturalism*, edited by Steven Gross, Tebben, Nicholas, and Williams, Michael, 180–209. Oxford: Oxford University Press.
Bar-On, Dorit, and Douglas C. Long. 2001. "Avowals and First-Person Privilege." *Philosophy and Phenomenological Research* 62: 311–335.
Bar-On, Dorit, and Douglas C. Long. 2003. "Knowing Selves: Expression, Truth, and Knowledge." In *Privileged Access: Philosophical Accounts of Self-Knowledge*, Ashgate Epistemology and Mind Series, edited by Brie Gertler, 179–212. Aldershot: Ashgate Publishing Limited.
Bar-On, Dorit, and Kate Nolfi. Forthcoming. *Belief Self-Knowledge*. Oxford Online Handbook.
Bilgrami, Akeel. 2006. *Self-Knowledge and Resentment*. Cambridge, MA: Harvard University Press.
Boyle, Matthew. 2010. "Bar-On on Self-Knowledge and Expression." *Acta Analytica* 25: 9–20.
Boyle, Matthew. 2011. "Transparent Self-Knowledge." *Aristotelian Society Supplementary* 85 (1): 223–241.
Byrne, Alex. 2005. "Introspection." *Philosophical Topics* 33 (1): 79–104.
Byrne, Alex. 2011. "Transparency, Belief, Intention." *Aristotelian Society Supplementary* 85 (1): 201–221.
Carruthers, Peter. 2011. *The Opacity of Mind: an Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.
Cassam, Quassim. 2014. *Self-Knowledge for Humans*. Oxford: Oxford University Press.
Coliva, Annalisa. 2012. "One Variety of Self-Knowledge: Constitutivism as Constructivism." In *The Self and Self-Knolwedge*, edited by Annalisa Coliva, 212–242. Oxford: Oxford University Press.
Evans, Gareth. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Fernández, Jordi. 2003. "Privilegesd Access Naturalized." *Philosophical Quarterly* 53: 352–372.
Gallois, André. 1996. *The Mind Within, the World Without*. Cambridge, UK: Cambridge University Press.
Gertler, Brie. 2011. *Self-Knowledge*. New York: Routledge.
McGeer, Victoria. 2004. "Autistic Self-Awareness." *Philosophy, Psychiatry and Psychology* 3 (2): 35–251.
Moran, Richard A. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press.
O'Shea, James R. 2007. *Wilfrid Sellars*. Cambridge, UK: Polity Press.
Nisbett, Richard, and Wilson, Timothy. 1977. "Telling More than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84: 231–259.
Rosenthal, David. 2005. *Consciousness and Mind*. Oxford: Oxford University Press.
Sellars, Wilfrid. 1956. "Empiricism and the Philosophy of Mind." In *Minnesota Studies in the Philosophy of Science*, edited by H. Feigl and M. Scriven Vol. I, 253–329. St. Paul: University of Minnesota Press.
Sellars, Wilfrid. 1969. "Language as Thought and as Communication." *Philosophy and Phenomenological Research* 29: 506–527.
Sellars, Wilfrid. 1975. "The Structure of Knowledge." In *Action, Knowledge and Reality: Studies in Honor of Wilfrid Sellars*, edited by Hector-Neri Castañeda, 295–347. Indianapolis: Bobbls-Merril.
Shoemaker, Sydney. 1968. "Self-Reference and Self-Awareness." *Journal of Philosophy* 65 (19): 555–567.
Wright, Crispin. 1998. "Self-Knowledge: The Wittgensteinian Legacy." In *Knowing Our Own Minds*, edited by C. Wright, B. Smith, and C. McDonald, 13–45. Oxford: Clarendon Press.