First-Person Authority: Dualism, Constitutivism, and Neo-Expressivism

Dorit Bar-On, UNC Chapel Hill

1. Introduction: Rorty's Dilemma

Nearly four decades ago, in a paper entitled "Incorrigibility as the Mark of the Mental", Richard Rorty argued that a genuinely *realist, non-eliminativist* view of mentality must provide a clear separation of the mental from the nonmental. Like Descartes, Rorty can be seen as maintaining that a proper separation requires understanding the special status of our spontaneous, non-evidential first-person self-ascriptions (in speech or in thought) of mental states – our so-called *avowals*.[1] And Rorty sides with Descartes in thinking that a proper separation will portray minded subjects as *incorrigible* in their avowals. Thus, a subject who spontaneously says sincerely, or thinks to herself: "I feel so tired", "I'm scared of this dog", "I hope I can finish this paper tonight" cannot be second-guessed. Others cannot be in a position to judge that, the subject's avowals notwithstanding, she is in fact *not* in the mental state she ascribes to herself. However, Rorty argued that such incorrigibility is inconsistent with a materialist conception of mental states. Thus, a committed materialist would be driven to *eliminativism*.[2]

What I shall call "Rorty's Dilemma" has us thus caught between the Scylla of Cartesian Dualism and the Charybdis of eliminativism about the mental. Proper recognition of what is distinctively mental requires accommodating incorrigibility, something Rorty thinks materialists cannot do. So we must either countenance mental states over and above physical states (and Cartesian Egos as proper subjects of mental states) in our ontology, or else give up altogether on the mental as a distinct category.

Consider first the eliminativist horn of the dilemma. Whatever the attraction of eliminativism in other areas may be, a thorough-going eliminativism about the mental seems to 'teeter into incoherence'[3]. For we would normally take someone who *claimed*: "There are no mental states" to be expressing her *conviction*, or *belief*, or *suspicion* or *impression*, etc., which are paradigmatic mental states. So there's a certain problem of uptake for the naïve interpreter of the eliminativist's claim. The eliminativist cannot avoid the difficulty by simply insisting: "There are no beliefs, but I don't believe that", for that comes too close to Moore's paradox. So there's a certain problem of delivery for the defender of the eliminativist claim. Finally, even setting these problems aside, it's not clear how the eliminativist could *explain away* the relevant trappings of mentalistic discourse. In other domains with suspect entities, the eliminativist can appeal to our illusions, mistaken beliefs, perceptions, ideas, concepts, fictions, metaphorical uses, or what have you, by way of explaining why we are led to conceive the world *as if* there are these suspect entities. But the eliminativist about the mental can hardly fall back on such mental substitutes in trying to explain away our mentalistic talk.[4]

Unlike eliminativism, Cartesian dualism doesn't appear to court incoherence, though it has notorious problems of its own. To fix ideas, let us discern the following four elements in the Cartesian conception of mind:

a. Absolute Epistemic Security: (What we've called) avowals enjoy absolute epistemic security– if a sincere, competent subject spontaneously self-ascribes being in a mental state M, then she cannot doubt that she's in M, her avowal is absolutly *infallible*, and is absolutely *incorrigible*;[5]

b. Privileged Access: Avowals represent occurrent self-beliefs that are obtained through the exercise of a special form of access that subjects of mental states have only to their own mental states;

c. Self-Intimation: The states that are known via this privileged form of access are by their nature *self-intimating* – if a subject is in a mental state M, she will inevitably believe that she is in M;

d. Substance Dualism: The states known via privileged access are states of an immaterial substance that is (only contingently) associated with our material body

As I read him, Rorty thinks that, if we embrace what are on their face *epistemic* marks of the mental offered by the Cartesian conception (viz. absolute epistemic security and privileged access), nothing short of a dualist ontology with self-intimating states will do. If so, then the above elements form a package from which substance dualism cannot be detached on pain of eliminativism. Interestingly, in generating his dilemma, Rorty does not place any weight on Self-Intimation. Note that the claim of Self-Intimation is different from the claim of Absolute Epistemic Security. Self-Intimation entails that *if a subject S is in mental state M, then she will inevitably think, or judge, or believe she's in M* (and thus self-ascribe M). By contrast, Absolute Epistemic Security commits us only to the claim that *if S spontaneously self-ascribes M, then her self-ascription is absolutely epistemically secure*.) Following Rorty, my initial focuse will be on the claim of Epistemic Security, though I will be working from the outset with a less hyperbolic version of it, dropping the requirement of *absolute* incorrigibility or infallibility. In this weaker form, I believe, the claim about avowals' security is much harder to deny than Self-Intimation.[6]

Now, Rorty's target in his article was the optimistic materialist of the day (he specifically discusses Smart and Armstrong), who thought – as many contemporary materialists continue to – that what is worth preserving in the Cartesian epistemic intuitions could be accommodated by *materialist introspectionism*. In section 2 below I briefly review some reasons for being dissatisfied with this materialist. In section 3, I outline two (by now familiar) *constitutivist* alternatives to materialist introspectionism. In section 4, I offer an alternative to both materialist introspectionism and constitutivism. This is the *neo-expressivist* view (developed and defended in Bar-On (2004)) according to which the distinctive status of avowals is to be explained by appeal to the expressive character of acts of issuing them (in speech *or* in thought). This view, I argue, allows us to stay clear of eliminativism without committing to Cartesian substance dualism, thereby offering an alternative way of slipping between the horns of Rorty's dilemma.

2. Materialist Introspectionism – Independence and Epistemic Authority

Although avowals resemble other everyday reports of contingent matters, such as "There is a cardinal at the bird-feeder", or "I have a mosquito bite on my leg", they appear to enjoy a peculiar epistemic status. In contrast to a host of other kinds of pronouncements (third-person present-tense mental ascriptions, past-tense mental self-ascriptions, present-tense self-attributions of psychological traits or standing dispositions, as well as non-observational bodily self-ascriptions and perceptual self-reports), avowals appear remarkably secure *even though* they are apparently made on *no* epistemic basis. Except under very special circumstances we take avowals completely at face value, we strongly presume them to be true, and we don't subject them to epistemic assessment or correction. When subjects *avow* being in a mental state (as opposed to making a mental self-ascription on the basis of self-observation or 'external' evidence, such as a therapy), the presumption is that they can 'just

tell', even though it's not obvious *how* they tell (nor does it typically seem appropriate to raise the question).

We can agree that avowals enjoy a distinctive security, even if we disagree with Rorty's hyperbolic characterization of them as "*absolutely* incorrigible", and even if we disagree on the *source* of their security.[7] Borrowing familiar terminology, in what follows I shall speak of the *first-person authority* that we enjoy when issuing avowals.[8] Now, according to *materialist introspectionism* ("MI" for short), first-person authority can be explained consistently with materialism by supposing that there are biological mechanisms – perhaps an intenral monitoring and scanning device – that allows us to obtain directly, or non-inferentially, information about the presence and character of these states. Although potentially fallible, corrigible, and dubitable, this mode of non-evidential 'privileged' access delivers highly reliable and accurate self-judgments that we articulate in our avowals.

We should separate two key elements in MI, which are in fact shared more broadly by a variety of views that fall under what I call the "epistemic approach":

> *Independence*: the thesis that mental states are physical states of our bodies whose existence and presence is ontologically (and conceptually) independent of anyone's beliefs or judgments *about* them; and
>
> *Epistemic Authority:* the thesis that first-person authority is a species of *epistemic* authority: the extent to which avowals are infallible or incorrigible is to be explained by appeal to the epistemic reliability or credibility of the method by which the self-judgments they articulate are arrived at.

In principle, opponents of MI's account of first-person authority can take issue with either Independence or Epistemic Authority (or both). In the next section, I will briefly

discuss accounts of first-person authority that sacrifice Independence. But right now I'd like to spell out one reason for questioning MI's version of the Epistemic Authority thesis. (My ultimate goal will be to persuade you that, if we want to find a way out of Rorty's dilemma, we should consider rejecting Epistemic Authority *without* jettisoning Independence.)

By identifying a distinct epistemic 'route' that we can associate with avowals, MI is able to accommodate a certain 1st-person/3rd-person *asymmetry* between avowals and mental ascriptions to others. But, as I understand him, Rorty thinks a proper account of first-person authority requires setting apart avowals from all *nonmental* self-ascriptions in terms of their security. Yet it looks as though MI would have difficulty drawing any principled distinction between avowals and a certain subset of bodily self-ascriptions – ones arrived at through proprioception, or kinesthesis, for example. After all, I have a different way from my observers of learning where my limbs are, or whether I am sitting down, but on any given occasion, my body could be in a state that my proprioceptive or kinesthetic mechanisms simply fail to distinguish from other states. In other words, I could be subject to a *brute local error* – an error that is not due to a defect in any of my psychological mechanisms but is simply due to the fact that my 'internal introspector' has been 'fooled' into false detection.[9]

It seems to me that, on the commonsense view, the possibility of brute local error is at least much more problematic than MI would have us expect. Commonsense clearly allows for false, corrigible avowals in cases of self-deception and wishful thinking. I myself think that under certain circumstances, even a sincere avowal of sensation may be thought false. For example, as you sit on the dentist's chair and say "My tooth hurts!" before the drill reaches your mouth, your dentist may sensibly question whether you really feel a toothache, even without questioning your sincerity.[10] But although these are cases in which

we are prepared to question the truth of an avowal, they do not provide examples of brute error; for these are precisely cases in which the avowal's falsity is assumed to be due to some psychological irregularity, failure, or defect on the part of the avower. We don't suppose in any of these cases that the subject might have *simply* been *fooled* into issuing a false avowal by an 'uncooperative' internal mental world.

As I see things, commonsense does not commit us to the absolute incorrigibility or infallibility of avowals. It does, however, seem committed to the *quasi-apriori* (albeit defeasible) *presumption* that avowals are true and are not to be subjected to ordinary epistemic correction and assessment. A philosophical view that aims properly to accommodate first person authority should explain what renders this presumption *reasonable* and identify the *source* of its quasi-apriori character. This is what MI fails to do. Thus, even if we relax Rorty's demands on what to expect of a proper demarcation of the mental in ways that are concessive to the materialist, it does not yet look like the materialist can discharge the explanatory burden thrown up by Rorty's dilemma.

Dissatisfaction with MI very often leads authors who are otherwise unsympathetic to Cartesianism to question the Independence thesis. Before turning to consider this option, however, I want to show how reflection on ordinary cases of false avowals may encourage us instead to reject Epistemic Authority. Consider again the above case of falsely (though sincerely) avowing feeling a toothache at the dentist. At first blush, this kind of case may seem grist to the materialist introspectionist mill: your tooth doesn't really hurt, but your internal introspector mistook what is in fact (say) fear of the approaching drill for pain, and it is this finding that your false avowal represents. But how plausible is it to regard your exclamation "My tooth hurts!" as a *report* on your condition, based on your (false) belief that your tooth hurts? After all, under the circumstances, you might have also said: "Ouch!", or

emitted a yelp, or winced. Should we think of the wince as equally the upshot of a *false belief* that you erroneously formed about your internal state? Isn't it much more plausible to regard the exclamation: "My tooth hurts!" as on a par with the wince – as something forced out of you, though in this unusual case not by an actual toothache, but rather by the priming effect of fear fueled by painful dental history?

In this somewhat roundabout way, we arrive at a well-known counter to introspectionism (in both its Cartesian and its materialist incarnations), which is often attributed to the later Wittgenstein. This is what I'll call *simple avowal expressivism*:: the view that, linguistic appearances to the contrary, avowals should be regarded as fundamentally similar to natural expressions such as grimaces, winces, smiles, and cries, in being bits of behavior that serve directly to express our present mental states rather than to report their presence. What is crucial about natural expressions is that they are in no way thought to articulate a subject's *judgment* or *belief* about some state of affairs. In particular, a subject's smile is not supposed to express *any* judgment, let alone her judgment *that* she is pleased; instead, it is supposed to give expression to the pleasure itself. Correlatively, we do not expect the person who smiles, or cries in pain, to give reasons for her smile or cry; we do not query or challenge a gasp of fear, or a sigh of relief, and so on. According to simple avowal expressivism, avowals are similarly protected by their 'logical grammar' from all epistemic challenge or assessment. And that's what accounts for their distinctive security.

Given our present concerns with questions of independence and irrealism, it may be useful briefly to compare simple avowal expressivism to a familiar expressivism in a different domain. According to traditional *ethical expressivism*, ethical claims (e.g., "Murder is wrong," "Racial discrimination is unjust") are mere expressions of certain of our emotions, preferences, or attitudes. They are not genuine, truth-evaluable assertions either about

objective ethical states of affairs (as ethical objectivists would have it) *or* about the expressed attitudes (as subjectivists would have it). It is useful to separate three different claims involved in traditional ethical expressivism:

(i) the positive *expressivist claim*: that ethical proclamations regularly serve to express pro- and con- attitudes, rather than to make assertions about objective states of affairs,

and

(ii) the negative *ontological* claim: that there are no ethical properties for ethical terms to refer to and no ethical facts for ethical sentences to describe,

as well as

(iii) the negative *semantic* claim: that ethical sentences are not truth-evaluable, since they do not express true or false propositions.[11]

Simple avowal expressivism clearly endorses an analogue of the positive expressivist claim. But notice that in this case the claim is offered as a way of explaining *epistemic asymmetries* – ones that arise not only between avowals and *non*-mental ascriptions but also between avowals and various *mental* ascriptions, including mental *self*-ascriptions. (An avowal: "I feel anxious" enjoys a greater and different security from a self-ascription with the same content made on the basis of, say, what your therapist told you.) It's not at all clear what sense to make of the analogue of the negative ontological claim in the case of avowals. If one had qualms about the reality of mental states one could hardly appeal to the fact that avowals directly express mental states in order to explain the asymmetries! Traditional ethical expressivism has aimed to preserve the cogency of ethical discourse even when it's denied that ethical claims are apt to describe genuine matters of fact. But if we take expressivism about psychological discourse to have a similar agenda, we court incoherence

(reminischent of the incoherence, pointed out earlier, that threatens eliminativism about the mental). Instead, avowal expressivism should be seen as attempting to explain epistemic asymmetries that arise in part *within* mentalistic discourse, without portraying first-person authority as a species of ordinary *epistemic* authority, underwritten by secure epistemic access.

What about the negative semantic claim? In the ethical case, we have the familiar "Frege-Geach problem": ethical sentences embed in a variety of force-stripping contexts, can partake in logical inferences, and so on, all of which seem to demand their possessing truth-conditions and being truth-evaluable. But the claim that avowals are just like natural expressions appears similarly to fly in the face of the *semantic continuity* between avowals and ordinary descriptive statements unlike natural expressions, avowals exhibit all the syntactic and semantic trappings of truth-evaluable self-ascriptions, such as "I am bleeding", or "I am walking down the street", including being contextually interchangeable with truth-evaluable statements (so, my avowal "I feel tired" seems truth-conditionally equivalent to "DB feels tired" which is ordinarily not issued as an expression of my feeling tired)[12].

Simple avowal expressivism does not directly take issue with the Independence thesis endoresed by MI. Instead, it rejects the Epistemic Authority thesis by denying that avowals represent judgments of *any* sort, and *ipso facto* denying that their security is due to the *epistemic* security of our reliably formed self-judgments. Insofar as this account purchases epistemic asymmetry at the cost of semantic continuity, however, I think it must be rejected. Where does that leave us?

3. Constitutivism

In his lectures on "Self-Knowledge and 'Inner Sense'", after developing an elaborate attack on materialist introspectionism and its kins, Sidney Shoemaker urges us to recognize

that "there is a conceptual, constitutive, connection between the existence of certain mental entities and their introspective accessibility" ((1994), p. 272), thereby rejecting Independence. A number of authors who, like Shoemaker, share the conviction that Epistemic Authority is to be denied, have similarly sought to deny the complete independence of mental states from subjects' judgments *about* them, by defending what has come to be called *constitutivism*: the idea that there are *constitutive connections* between the nature and presence of first-order mental states and the presence of correct self-ascriptive judgments.[13]

In what follows, I'll briefly canvass two main varieties of constitutivism. On the first, 'grammatical' variety, the category of the mental is to be delineated through 'logical grammar' and recognition of first-person authority is simply a 'bedrock' condition on mastery of the rules of mentalistic discourse. On this variety, the assumption of first-person authority is a matter of the conventions governing mentalistic discourse. On the second, 'ontological' latter variety, by contrast, the assumption of first-person authority is a direct reflection of our apprehension of the nature of mental states, which in turn determines how the category of the mental is to be delineated.

In several places, Crispin Wright considers a non-expressivist alternative to epistemic views of first-person authority, which he dubs "the default view". On this view, mentalistic discourse is governed by "a *constitutive principle*" that "enters primitively into the conditions of identification" of the subject's mental states (Wright 1991: 142). The principle constrains the truth-conditions of mentalistic ascriptions as follows: "unless you can show how to make better sense of her by overriding or going beyond it, [a subject's] active self-conception, as manifest in what she is willing to avow, must be deferred to" (Wright 1998: 41). Unlike simple avowal expressivism, the default view can accommodate the semantic continuities between mental ascriptions to oneself and to others; in particular, it does not require denying

truth-evaluability to avowals. Still, insofar as it presents the epistemic asymmetries between avowals and other ascriptions as simply due to the conventions governing mentalistic discourse, there is a clear sense in which the default view, like simple avowal expressivism, is a 'grammtical' view.[14]

In some places, Wright offers a more ontological articulation of the default view, using a kind of anti-realist model from other areas of discourse — e.g., color discourse, and, on some views, ethical discourse. In these areas, it has seemed misguided to some philosophers to conceive of successful judgments as tracking a completely mind-independent reality; so it is argued that the alleged facts exhibit *judgment-dependence*. Thus, what is true about an object's color, for example, is said to be systematically dependent on the color-judgments of well-placed perceivers.[15] In the case at hand—the mental realm—the analogous claim would be that whether, e.g., the avowal "I feel awful" is true or not, and whether or not the self-ascriber does feel awful, is not entirely independent of what she *thinks* about her state. Since this dependence on the subject's own verdict constrains any ascription of a mental state to her, if someone said of me: "She feels awful", the truth of her ascription would presumably also not be independent of what *I* take to be the case about my state. If so, it should be no surprise that there is such a fit between what a subject says (or thinks) about her condition and the truth of the matter, and ultimately no point in seeking an explanation of the asymmetries between avowals and other ascriptions in terms of special epistemic access.

However, the mentalistic case seems to me different from paradigm cases of judgment-dependence. In paradigm cases, the idea of judgment-dependence is invoked to capture a contrast between one range of facts (concerning, e.g., secondary qualities) and other sorts of 'fully objective' facts, where the intuition is that even an ideally well-placed

judge could go wrong in her judgments. By contrast, the phenomenon of first-person authority marks a partition *within* the mental realm; it doesn't merely represent a contrast (in terms of degree of success) between our judgments in the mental realm *as a whole* and our judgments in other areas. The intuition is that subjects are better placed than their observers with respect to a single range of facts, but one's being well-placed to pass a 'truth-determining' judgment requires that the state be one's own, that the judgment be concurrent with the state's presence, and that it be issued in the *'avowing mode'* (i.e., without reliance on ordinary observation, evidence, or inference). This means that, in the mental case, unlike in the case of secondary qualities, our understanding of the range of facts to be construed as judgment-dependent depends on our understanding of what it is for someone to be well-placed to pass judgments on them. And this seems to cast doubt on the analogy to the color case and undermine the explanatory usefulness of the judgment-dependence model.

Another, perhaps related, difference comes to mind. In the color case, the idea is that we are to construe colors, which *appear* to be properties of mind-independent objects, as *in reality*, dependent on visual judgments or responses of color perceivers. It is natural to understand this idea in a *reductive, irrealist* way. But such a reading is problematic in the mental case, for reasons directly related to Wright's own worries about wholesale psychological anti-realism or eliminativism (mentioned breifly in my opening section). Wright thinks that the psychological anti-realist who models her view after expressivism in ethics or after eliminativism in other areas "teeters into incoherence", because her "very claim … presupposes an underpinning in facts about aspects of the characteristic attitudinal psychology of its participants. But facts of that genre are just what radical anti-realism about psychology is unwilling to countenance".[16] On the face of it, however, applying the model of judgment-dependence in the mental case faces a similar difficulty. For construing mental facts as

judgment-dependent would seem to require the invocation of the very same kind of states that the judgment-dependence theorist is trying to reconstruct.

Some ontological versions of constitutivism avoid trading on the idea that mental self-judgments are extension-determining by maintaining that (at least in creatures like us), the belief that one is in a mental state constitutes part of the metaphysical identity conditions of the mental state.[17] The idea is that genuine mental states depend for their very identity (and thus presence) on being self-ascribed by their 'hosts', so that I couldn't count as e.g., wishing that it doesn't rain tomorrow, in the first place, unless I also thought that I was so wishing. A familiar complaint about this idea is that it rules out too much. On this ontological constitutivist view, we could not bring under the umbrella of the mental unconscious emotions, wishes, and thoughts, unnoticed or unattended-to sensations and feelings, regardless of how similar these states appear to be to the allegedly genuine mental states in terms of the behavioral dispositions they issued in. Moreover, we'd also have to treat as a separate (sub-)category those psychological states that we share with infants and non-human animals. To make their thesis more palatable, constitutivists typically restrict it to *rational* beliefs and intentions, *'judgment-sensitive'* wishes, desires, and preferences, intentional states understood as *commitments*, and so on. And they suggest that it is only these states that belong in the category of the Mental properly so-called. Other states that are often labelled 'mental' – brute sensations, passing thoughts and perceptions, intrusive cravings and impulsive wants or irrational beliefs, and so on – would all need to be relegated to a second class, 'lower case' mental status. The result is a view that may be described as "Mental-mental dualism".[18]

I cannot here undertake a fair presentation or critique of this view. For present purposes it suffices to make the following observations. First, insofar as the ontological constitutivist invokes a key (and controversial) element in the Cartesian 'package', namely,

Self-Intimation, and inasmuch as it is driven to a Mental-mental dualism, it isn't clear how it can be said to escape Rorty's dilemma. I suspect that at least some of the reasons Rorty has for doubting that materialism could accommodate incorrigibility would lead him equally to doubt that it could accommodate Self-Intimation and make room for a separate category of the Mental. But, second, even setting aside these reasons, it is clear that the constitutivist account can at best explain the security of avowals of Mental states – since the constitutivist thesis is supposed to apply only to them. I think we should agree with Rorty that avowals' security extends beyond that narrower range. We seem to enjoy first-person authority when pronouncing on our passing thoughts ("I'm thinking that there's water in this glass"), whims ("I just feel like jumping"), sensations ("I'm feeling nauteous"), irrational convictions and wants ("I feel confident so-&-so will win", and so on. This apparent authority, whether or not it is of a piece with the kind of authority we enjoy over our more considered judgments, intentions, demands explanation. Yet it remains untouched by the constitutivist's account.[19]

Setting aside eliminativism, suppose we are dissatisfied with simple expressivism, as well as with introspectionist accounts of first-person authority. Then, given the foregoing discussion, it looks like we may be caught in a descendant of Rorty's dilemma. We must *either* surrender the mental/nonmental divide to grammar/convention, and deny it any metaphysical substance, *or* we adopt a metaphysical divide that rests on a new form of dualism (albeit one that doesn't commit us to immaterial substances), namely Mental-mental dualism, and settle for a rather limited account of first-person authority.

4. A Neo-Expressivist View[20]

In the remainder of this paper I want to explain how my preferred, *neo-expressivist* account of avowals' security attempts to capture and explain the phenomenon that gives rise

to Rorty's original dilemma and its descendent. Like constitutivism, the neo-expressivist account does not explain first-person authority by appeal to the Epistemic Authority thesis, but unlike constituvism, it does not require rejecting the Independence thesis and it allows that we enjoy first-person authority in all our avowals, not only those of Mental states.[21]

The neo-expressivist account takes its lead from an account we rejected earlier: simple avowal expressivism. What seems promising about that account is the positive *expressivist claim*. But, to avoid compromising the semantic continuities between avowals and other ascriptions, the expressivist claim must be decoupled from the negative semantic claim.[22] And, to avoid landing in an irrealism that is only dubiously coherent, it must also shy away from the negative ontological claim.

As a preliminary, let us go back to self-ascriptions issued through proprioception or kinesthethesis. These self-ascriptions surely represent genuine truth-evaluable reports or judgments; but they share a certain epistemic feature with avowals. Proprioceptive and kinesthetic self-reports are "identification-free": epistemically speaking, they do not rest on a recognitional identification of the subject of the utterance or thought. That is to say, in normal circumstances, if I say (or think): "My legs are crossed" or "I'm spinning around", my utterance/thought does not rest on my *recognizing* some individual as myself. I do not identify someone as being me and take that person's legs to be crossed. And I have no more resason for thinking that *someone*'s legs are crossed than whatever reason I have for thinking that *my* legs are crossed. Bodily self-reports of this kind are "immune to error through misidentification" (IETM, for short).[23]

When a self-ascription of the form "I am F" is IETM, then, although I may fail to be F, so my self-ascription may be false, there is no room for me to think: Someone is F, but is it *me*? This is because, *on that occasion*, I have no specific reason for thinking that *someone* has

the relevant properties over and above, or separately from, whatever reason I have for thinking that *I* have them. (Contrast this with a case in which I, e.g., tell how much money I have in my bank account by consulting the bank teller's screen. Here, the information I receive that *someone* has (say) $500 in her account, coupled with my taking that information to be about *my* account, gives me reason to think that it is *me* who has $500.) Note that immunity to error through misascription does not reflect a special recognitional success, i.e., success in identifying – in the sense of singling out – the "right" individual of whom to predicate F. Quite the opposite: if anything, this kind of immunity reflects the *absence* of recognitional identification.[24]

Now, on Evans' analysis, self-ascriptions that are IETM can still represent *knowledge* that we gain about ourselves in a distinctive way. Evans notes that we possess two general capacities for gaining information about some of *our own* states and properties – "a general capacity to perceive our own bodies" (which includes "our proprioceptive sense, our sense of balance, of heat and cold, and pressure"), and 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."[25] When a subject gains information of the form "I am F" (for the relevant range of F's) in one of these ways, Evans remarks, "[t]here just does not appear to be a gap between the subject's having information (or appearing to have information), in the appropriate way, that the property of being F is instantiated, and his having information (or appearing to have information that that *he* is F; for him to have, or to appear to have, the information that the property is instantiated just is for it to appear to him that *he* is F." ((1982), p. 221) But, for all that, self-ascriptions of the form "I am F" that are IETM can still represent secure knowledge I have that *I* myself am F.

My neo-expressivist account begins with the suggestion that ordinary present-tense mental self-ascriptions of the form "I (am) M(ing) (*that) c*," where M is a mental state and *c* is its putative intentional content,[26] *when issued as avowals*, are *immune to error through misascription*, in addition to being IETM. Suppose I avow: "I am nervous about this dog," or "I am

hoping that you'll join us tonight", or "I am feeling tired". My self-ascription, I maintain, does not reflect a judgment I pass upon in(tro)specting the appearances of an internal state of mine with an "inner eye" or an internal scanner. When avowing, my self-ascription does not rest on my recognition of such an item as falling under the category M (as opposed to M') and having content *c* (rather than *c'*). If I *avow* being in M, I have no reason – specific to the occasion on which I issue the avowal[27] – for thinking that I am in *some* state or other, and that it has some content, other than whatever reason I have for thinking that the state that I am in is mental state M and its content is *c*.[28]

To say that avowals are immune to error through misascription is not to say that they are absolutely infallible or incorrigible. It's just to say that they are protected from a certain array of epistemic errors (and thus corrections) – a *much wider* array than other ascriptions, *including*, specifically, proprioceptive and kinesthetic self-reports, as well as evidential mental self-ascriptions. If I say or think: "My legs are crossed" (in the normal way), then even if I cannot be misidentifying who it is whose legs are crossed, if my legs are *not* crossed, my self-report will be mistaken precisely because I have misidentified the state of my legs: I will have mistaken one state of my limbs for another. But in the case of avowals, I suggest, though my self-ascription can be mistak*en,* if it is, this is not because it involves some recognitional mis-tak*ing* of the state I am in. Thus, if I avow feeling scared of the dog, and as it turns out, the creature in front of me isn't a dog, but (say) a coyote, my avowal will be false, but my failure in such a case will not be due to my *taking* my state to have the wrong intentional object. (After all, under the circumstances, I would be equally disposed to say or think non-self-ascriptively "That dog is scary!"). Similarly for cases in which a subject avows feeling annoyed at someone when in fact (we might think) he is actually feeling attracted to her. I would argue that, in the relevant circumstances, it is no more plausible to regard the avowal "I find her annoying" as grounded in one's mistaken recognition of one's state as one of annoynace than it would be to so regard the non- self-ascriptive claim "She's so annoying". As we saw, in certain circumstances we may even be led to question an avowal of pain. But

I would argue that these are not circumstances in which the avower mistakes some other state of hers for pain due to the way the state *appears* to her. After all, in the circumstances, she would be equally disposed to emit a spontaneous yelp. (And, on pain of regress, it is implausible to suggest that the yelp would be the result of a false *self-belief* based on how her present state *appears* to her.)

The characterization of avowals' distinctive security as a matter of their unique immunity to error through misascription provides a suitably tempered interpretation of the claim that avowals are incorrigible, which does not require invoking either Cartesian privileged access or, indeed, any distinctively secure epistemic basis on which avowals supposedly rest. However, by itself, the characterization does not give us a full explanation of avowals' special security. For one thing, we need to understand the *source* of the additional immunity to error that avowals enjoy.[29] *Why* is it that avowals are not only immune to error through misidentification but are also immune to error through misascription? In addition, we have seen that immunity to error, in general, is no guarantee of *truth*. So we need to understand why avowals contrast with other ascriptions in being governed by a quasi apriori presumption of truth.

It is at this point that I think we should co-opt a key insight from simple avowal expressivism: the idea that avowals' security is to be explained by appeal to their *expressive character*, rather than by appeal to the security of this or that epistemic basis on which they are made. Like simple avowal expressivism, *avowal neo-expressivism* maintains that the epistemic simplicity and immediacy of avowals is best explained by seeing them as direct *expressions of subjects' self-ascribed mental states*. But in contrast to simple avowal expressivism, neo-expressivism does not rely exclusively on the idea that avowals are just like inarticulate

grunts, grimaces, or cries, and maintains that avowals, like various mental and non-mental reports represent genuine, truth-evaluable self-ascriptions.[30]

Semantically speaking, a self-ascription that is IETM, such as "I am sitting down" (as normally made) is *about* a particular individual in the world, me. But, epistemically speaking, it does not have as its ground or reason a recognitional identification of myself as the one who is sitting down. If it is epistemically warranted, this is because in issuing the self-ascription I put to use certain distinctive capacities for gaining information about the relevant state of affairs (viz., my sitting down) – capacities that I have only with respect to certain of my own states. I wish to make an analogous separation regarding self-ascriptions that enjoy, as I put it, immune to error through misascription. Semantically speaking, an avowal such as "I am hoping that dinner will be served soon" ascribes to me a present state of mind, with a particular content. However, the avowal does not have as its reason the recognitional judgment to the effect that it is a state of hoping that I am in hope, or that it is the content *that dinner will be served soon* that my state has. (So the avowal is not grounded in recognitional jugments concerning *either* the identity of the state's bearer *or* the presence and content of the particular state in herself.) If it is epistemically warranted, this would be because, in issuing it, I put to use a distinctive capacity that I have only with respect to certain of my own states. The capacity in question, I propose, is the *expressive capacity to give articulate voice to my present states of mind, or to speak from them* – a capacity only I have, and only with respect to my mental states, and one that I put to use only when *avowing* my mental states, as opposed to reporting my findings about them on this or that basis.

In what follows, I briefly introduce some distinctions that would allow us to give more precise sense to the neo-expressivist claims. I will then return to the question of the neo-expressivist explanation of first-person authority.

The first distinction is one borrowed from Wilfred Sellars (1969), between two different senses (or kinds) of expression.

> *a*-expression**:** in the *action* sense, a *person* expresses a *state* of hers by intentionally doing something;

For example, when I give you a hug, or say: "It's so great to see you," I express in the action sense my joy at seeing you. One may also express one's feeling of sadness in the action sense by *letting* tears roll down her cheeks, instead of wiping them out and collecting oneself. Note that the notion of a-expression requires that a person do *something* intentionally. It does not require that what one does intentionally is *express*. One *can* intentionally express a mental state – for example, by deciding or setting out to give vent to a present emotion, instead of suppressing it. But the more basic case is one in which a person gives *spontaneous* expression to a present state of hers through performing some intentional act that doesn't have expression as its intentional aim or purpose.[31] Note that one can a-express a mental state in thought, and not just in speech. Annoyed at a friend, you may silently curse, or say to yourself "I hate you!"

> *s*-expression: in the *semantic* sense, e.g., a *sentence* expresses an abstract *proposition*, thought or judgment by being a (conventional) representation of it.

For example, the sentence "It's raining outside" expresses in the semantic sense the proposition that it is raining at time [*t*] outside place [*p*]. The notion of s-expression, too, is applicable to thought as well as speech. We can speak of a thought-token as expressing a proposition, and we can extend the relation of s-expression to cover unstructured linguistic expressions (e.g., predicates are said to express general concepts), and their analogues in thought. The important thing to keep in mind is that s-expression is a semantic relation,

holding between linguistic expressions (and their analogues in thought) and their meanings, whereas a-expression is a relation holding between persons (and other agents) and mental states.[32]

As we consider the respects in which avowals resemble natural expressions, we should be focusing on natural expressions that fall under a-expression – ones produced in the course of performing intentional actions (e.g., giving a smile or a hug, or letting out a sigh). Furthermore, as we consider these cases, we must distinguish between the *act* of expressing and its *product*. Like other nominals in English (and other languages), such as "statement," "assertion," "ascription," "report" (as well as "expression") etc., "avowal" is ambiguous between the act of avow*ing*, which is an event in the world with a certain causal history and certain action-properties, and the result or product of such an act—a linguistic (or language-like) token, an item with certain semantic properties. The product of an act of avowing, unlike a smile or a wince, or even a verbal cry such as "Ouch!", is a semantically articulate self-ascription with semantic structure and truth-conditions. Avowals understood as products s-express self-ascriptive propositions, to the effect that the avower is in some state. Natural expressions such as a facial expression, a gesture, a bodily movement, a demeanor, an inarticulate sound, when understood as products, and however produced—do not on their face s-express anything. There are no semantic conventions in virtue of which laughter refers to amusement, no linguistic rules in virtue of which a hug signifies, or stands for, joy at seeing the person hugged.[33] Thus there are notable differences between avowals and acts of natural expression in terms of their products. (I would argue, however, that we should recognize important continuities between natural expressions and linguistic expressions even when understood as products.[34]) But I think that the expressivist

insight regarding avowals should be understood, in the first instance, as a claim about the relevant *acts*, not about their products. The claim is that there are notable similarities between the act of avowing a state and the act of giving it a natural expression; e.g., between the act of saying out loud or to oneself "I hate this mess!" (or, alternatively, "This is such a mess!" or "Ugh!"), and sighing in exasperation, stomping one's foot, etc. And this claim could be true, even if there were systematic differences between the products of acts of avowing and the products of naturally expressive acts.

Expressive acts are typically not acts performed with a prior intention or with a specific goal or purpose in mind. Nonetheless, such acts seem to meet one plausible characterization of intentional acts (due to Anscombe (1957)). They are voluntarily produced bits of behavior, where the person producing them *knows what she's doing* and where that knowledge is *nonobservational*. If our ordinary reason-giving practices are any indication, our reasons for such doings are typically *not* beliefs or thoughts we have *about* the expressed mental states but rather simply our being in the states: I'm giving a hug because I feel happy to see you; I'm sighing because I feel exasperated. But our avowals, I submit, can be performed for the very same sorts of reasons: when I avow "I am so happy to see you" I say this precisely because *I feel so happy to see you*.

So avowals, on my neo-expressivist view, constitute a certain class of *expressive acts* in which a subject gives articulate vent, in speech or in thought, to present mental states. When performed out loud, these are acts of *speaking one's mind self-ascriptively*, instead of giving either non-linguistic or else non-self-ascriptive expression to it. Such acts are epistemically unmediated, even though they issue in true or false products (sentence or thought tokens). *Like* simple avowal expressivism, the neo-expressivist view does not explain avowals' distinctive features by appeal to their secure epistemic

basis, but instead appeals to their expressive character. However, unlike the simple view, it does not take the security to be a matter of the *semantics* or 'grammar' of avowals understood as products; it does not require denying that avowals, *qua* products, are genuine self-ascriptions that can share truth-conditions with other ascriptions (e.g. "I am tired" avowed by me is true iff DB is tired). The neo-expressivist positive claim is confined to what avowals a-express and does not concern what avowals (*qua* products) s-express. Thus, it does not get coupled with a negative semantic claim (viz., a denial of truth-conditions and thus of truth-evaluability).

As for first-person authority, the neo-expressivist account does not take it to be the epistemic authority of an expert, nor does it portray it as a feature that is simply built into the 'grammar' of mentalistic discourse by pure convention *or* into the very nature of mental states. Rather, on the neo-expressivist account, first-person authority reflects our recognition of subjects' avowals as acts in which they ascribe mental states to themselves at the same time as they give them direct expression. *When avowing, I a-expresses the very same state that my avowal as product says I am in. Thus, to take me to be avowing is to take it that I am in the very state whose presence is required to make my avowal true.* This is what explains the quasi-apriority as well as reasonableness of the presumptions that comprise so-called first-person authority, as well as its inalienable and non-transferrable character.

On constitutivist views of the ontological variety, *being* in a mental state is partially constituted by thinking that one is in the mental state, which means that we must deny Independence. By contrast, neo-expressivism does not require endorsing claims about necessary, constitutive connections between the existence and presence of mental states and subjects' judgments. The view allows for a wide array of errors and failures of self-judgment

– the same array, I would argue, that is allowed by commonsense. In general, it allows that being in a mental state is one thing; passing judgment that one is in the mental state is another. Except in special cases, judging that one is in a mental state does not make it so, and one can be in a mental state without judging that one is. Moreover, the view even allows that one can be in a mental state without showing it through expressive behavior, and one can engage in behavior expressive of, say, joy, without expressing *her* joy (as when one merely 'puts on' a happy face). By the same token, one can be in a mental state without avowing it, even in thought, and, I've suggested, one can also issue a false avowal. Thus, in the general case, our avowals are answerable to the facts about our mental lives, just like others' (or our own) reports about them.

Now what about Rorty's dilemma and its descendent? If the neo-expressivist account is right, we may have the materials for recovering the commonsense separation of mind and body *without* resorting to Cartesian dualism but also without a wholesale rejection of Independence. Far from committing us to an ontology of hidden states that are only directly observable by subjects and inferrable by their observers, the commonsense picture that neo-expressivism tries to articulate is one according to which, as observers, we can normally tell what mental states others are in because they are shown in their expressive behavior (avowals included).[35] What subjects can do that others who observe them cannot is express their states of mind either through non-verbal behavior or by *speaking from* them. If anything, *expressibility,* and not incorrigibility, is the mark of the mental. Mental states, on this view, are states that are *expressible*, though they need not always be express*ed*. However, *this* dependence – of mentality on expressibility – is not the dependence endorsed by constitutivism, of the nature, existence, or presence of mental states on mental self-judgments. For all that, I have argued, the neo-expressivist can still account for first-person

authority, now to be understood neither as a matter of the epistemic reliability of our higher-order judgments nor as something built into the grammar of mentalistic discourse or the very nature of mental states, but rather as the privilege of speaking one's mind.

Acknowledgements

I wish to thank Matthew Boyle and Ted Parent, as well as two anonymous referees, for helpful comments on earlier drafts and to Ted Parent for much-appreciated editorial work. Earlier incarnations of the paper were presented at the "First-Person Authority" conference in Duisburg, Germany (Sept 2007), the Auburn University Colloquium (Oct 2007), and the "Self-Knowledge and the Self" conference in London, England (May 2008); I thank the audiences at these talks for helpful discussions. I have also benefited from written and oral exchanges with Quassim Cassam, Eric Marcus, Ram Neta, Jim Pryor, Sidney Shoemaker, and Keith Simmons.

---

[1] As I shall use the term, avowals can be made in speech *or* in thought. I can say *or* think in the distinctive first-person way "I am sick and tired of this mess".

[2] Rorty himself had already embraced eliminativism as early as in his (1965); see also his (1970b).

[3] I here borrow a phrase from Wright (1995).

[4] For defense of psychological eliminativism, see Stich (1983). For discussion, see Hannan (1993).

[5] Rorty (1970a) focuses on absolute incorrigibility, since he thinks that infallibility and indubitability (as well as privacy) are not feasible candidates for marks of the mental.

[6] I will return to Self-Intimation later on.

As will become clear, I'm using Rorty's dilemma as a foil for making vivid a certain explanatory challenge: to offer a genuine explanation of so-called first-person authority, while staying clear of Cartesian dualism, consistently with materialist ontology.

[7] For some congenial characterizations of the phenomenon of interest, see e.g., Wright (1998), Moran, (2001), Heal (2002), Bilgrami (1998), and Smith (1998). I shy away from characterizing avowals' epistemic security in terms of privileged *knowledge* for reasons I explain in (2004), ch. 1 and *passim*.

[8] See Davidson (1984), where the phrase "first-person authority" is apparently first coined.

[9] Other types of errors allowed by MI are discussed in Bar-On (2004), ch. 4. For relevant discussion, see Wright (1998) and Heal (2002).

[10] I discuss other examples of sincere but false avowals of sensations in (2004), pp. 329-335, 394-396. I thus disagree with Wright that avowals of sensations ('phenomenal' avowals) differ from attitudinal avowals in being 'strongly authoritative' (see Wright (1998), p. 14).

[11] The closely related (but still separate) ethical *noncognitivist* claim is that ethical sentences never express cognitive, truth-evaluable beliefs or judgments.

[12] Just as one can issue a genuine self-report using the first-person "I" (as in saying or thinking "I'm depressed", based on what your therapist has convinced you of), one can express a present mental state using a self-asctiption that doesn't use "I" (as when a parent says to a child: "Daddy would like you to eat now").

[13] See, e.g., Boghossian (1997), Wright (1998), Bilgrami (1998), Moran (2001), Heal (2002). For reasons of space, I cannot here undertake a detailed discussion of the various versions of constitutivism defended by the different authors.

[14] (Wright 1998: 41). The discussion here follows closely my discussion in (Bar-On (2004), pp. 347-350, 410-412). For other discussions of the default view, see Boghossian (1989), McDowell (1998), Fricker (1998), Bilgrami (1998), and Moran (2001). Zimmerman (2006) describes it as 'anti-realist constitutivism' and contrasts it with Shomeaker's 'realist constitutivism'.

[15] For discussion of this idea, see e.g. Wright (1989), sec. 3, and (1992), ch. 3, appendix.

[16] Wright (2002), p. 213; see also Wright (1995), pp. 204-205.

[17] Shoemaker (1996) suggests that self-belief is part of the 'functional profile' of (at least some) mental states. This functionalist characterization is closer to the constitutivism that Shoemaker prefers over the more causal (even if lawlike) generalization that, in creatures like us, "one's being in a certain mental state produces in one … the belief that one is in that mental state" (225).

[18] See Moran (2001), Heal (2002), Bilgrami (2006). See also Coliva (2006).

[19] In response to the objection that the constitutivist strategy is too restrictive, Bilgrami (2006) advocates admitting from the start that Self-Intimation fails in many cases of putative mentality (e.g., higher animals and infants), but treating these as cases in which the relevant subject simply fails to be in a *genuinely* mental (=Mental) state. (For a brief discussion of Bilgrami's strategy, see Bar-On (2007); a proper treatment will have to await another occasion.) It's not clear how this strategy will fare any better in connection with Rorty's dilemma. Rorty himself also separates a subcategory of the properly mental states within the broader category of the psychological; see his (1970a). However, his category of the mental covers many types of states that Bilgrami and other constitutivists would exclude from their preferred Mental states.

[20] This section contains a very brief summary of key ideas developed in my (2004), esp. chapters 4-8.

[21] To clarify, I will not be concerned to deny that there's a *species* of first-person authority associated with a certain subset of our avowals that is best captured by a constitutivist account. However, such an account will leave the broader phenomenon untouched, and it will remain to be seen whether a materialist could accommodate it.

[22] Compare Wright (1988), p. 30. An anonymous referee has pointed out that E. Tugendhat proposes a combination of epistemic asymmetry and semantic continuity in *Self-Consciousness and Self-Determination* (English translation: Cambridge, Mass./London: MIT Press, 1986).

[23] For discussion of the phenomenon of immunity to error through misidentification, see Wittgenstein (1958), pp. 66-67, Evans (1982) (esp. ch. 7, sec. 2), Shoemaker (1968), and Wright (1998), pp. 18-20.

[24] There is a "thin" sense in which I do identify myself as the subject of the ascription. It may be useful here to distinguish between the referential notion of identifying (a semantic notion) and the recognitional notion of identifying (an epistemic notion). Compare Evans (1982), p. 218.

[25] Evans (1982), pp. 220, 222.

[26] I am using the phrase "intentional content" to cover both intentional object (e.g. "I'm afraid of *the dog*") and propositional content (e.g. "I'm hoping that *it won't rain today*").

[27] I say this to rule out general inductive evidence I may have for thinking that I am in some mental state or other. (Thanks to Sydney Shoemaker, Jim Pryor, and Ram Neta.)

[28] When ascribing mental states to others, I *do* typically have such independent reasons. Noticing that you're scared of *something*, I may need to figure out what it is you are scared *of*, and conjecture that it's the dog. But I can sensibly wonder whether it is the cat instead, even as I continue to be convinced you're scared of *something*.

(Under special circumstances, I may also wonder what *I* am scared of – but not, I'd argue, when I'm simply avowing being scared of x.)

[29] The immunity to error through misidentification of proprioceptive reports has to do with our possessing special mechanisms for obtaining information concerning our *own* bodies. See Evans (1982), ch. 7.

[30] In (2004), Ch. 6, I motivate neo-expressivism first by considering thought avowals, such as "I am thinking that there's water in the glass", understood as a self-ascription of a presently entertained thought. I argue that their self-verifying character (see Burge (1988)) is to be explained by the fact that the avower is *articulating* the very content she assigns to her thought. I then offer a neo-expressivist construal of this idea and extend it to the assignment of content to other mental states, as well as to the mental state component.

[31] Rosalind Hurstouse (1991) provides examples of expressive acts precisely to undermine the received dogma that only behavior backed up by a Davidsonian belief-desire pair can be regarded as intentional. See also Green (2007). The key point is that just because avowals have semantically articulate, self-ascriptive products they need not have belief-desire pairs as their reasons.

[32] See Sellars (1969), pp. 506-27, where he also mentions expressing in the *causal* sense, which I here set aside.

[33] Some gestures and other non-verbal expressions are governed by socio-cultural conventions, but the conventions do not assign semantic content; instead they set up a ('pragmatic') connection between *making* the gesture (e.g., tipping one's hat) and being in the relevant state, or having the relevant attitude, sentiment, etc. (e.g. feeling respect). On the other hand, an animal's alarm call may be thought to have semantic content – representing e.g. *threat from above*. Still, unlike an avowal, it cannot be plausibly taken to express in the semantic sense a proposition *about* the relevant mental state of the animal issuing the call.

[34] Such continuities can be exploited in trying to understand how so-called nonnatural meaning could arise in a world where natural meaning is to be found. (See Bar-On and Green (in progress).)

[35] For more on this, see Bar-On (2004), pp. 410ff.

References

Anscombe, E. (1957). *Intention*. (Cambridge, MA: Harvard University Press)

Bar-On, D. (2004). *Speaking my mind: Expression and self-knowledge*. (Oxford: Oxford University Press)

_____ (2007). Review of Akeel Bilgrami, *Self-knowledge and resentment*. *Notre Dame Philosophical Reviews* (September)

Bar-On, D. & Green, M. (in progress). Expression, communication, and meaning.

Bilgrami, A. (1998). resentment and self-knowledge. (In Wright et al. (1998), 207-241.)

_____ (2006). *Self-knowledge and resentment*. (Cambridge, MA & London: Harvard University Press).

Boghossian, P. (1989). Content and self-knowledge. *Philosophical Topics*, *17*, 5-26

_____ (1997). What the externalist can know a priori. (In Wright et al. (1998), 271-284. (originally published in *Proceedings of the Aristotelian Society*, *97* (1997), 161-175)

Burge, T. (1988). Individualism and self-knowledge. *Journal of Philosophy*, *85*, 649-663

Coliva, A. (2008). Self-knowledge: One more constitutive view. *Synthese*.

Davidson, D. (1984). First person authority. *Dialectica*, *38*, 101-111

Fricker, E. (1998). Self-knowledge: Special access vs. artefact of grammar–A dichtomy rejected. (In Wright et al. (1998), 155—206)

Green, M. (2007). *Self-expression*. (Oxford University Press)

Hannan, B. (1993). Don't stop believing: The case against eliminative materialism. *Mind and language*, *8*, 165-179

Heal, J. (2002). *Mind, reason, and imagination*. (Cambridge: Cambridge University Press)

Hursthouse, R. (1991). Arational actions. *Journal of Philosophy*, *88*

Lycan, W. G. (unpublished manuscript). *The superiority of HOP to HOT*. Retrieved from http://www.unc.edu/ ~ujanel/HOPHOT.htm

McDowell, J. (1998). Response to Crispin Wright. (In Wright et al. (1998), 47-62)

Moran, R. (2001). *Authority and enstrangement: An essay on self-knowledge*. (Princeton, NJ: Princeton University Press)

Peacocke, C. (1998). Our entitlement to self-knowledge: Entitlement, self-knowledge, and conceptual redeployment. *Proceedings of the Aristotelian Society*, *96*, 117-158

Rorty, R. (1965). Mind-body identity, privacy, and categories. *Review of Metaphysics*, *12*, 24-54

_____ (1970a). Incorrigibility as the mark of the mental. *Journal of Philosophy*, *12*, 399-424

_____ (1970b). In defense of eliminative materialism. *Review of Metaphysics*, *24*, 112-121

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*, 329-359

Shoemaker, S. (1968). Self-reference and self-awareness. *Journal of Philosophy*, *65*, 555-567

_____ (1994). Self-knowledge and "inner sense". *Philosophy and Phenomenological Research*, *54*, 249-314

Sellars, W. (1969). Language as thought and as communication. *Philosophy and Phenomenological Research*, *11*, 506-527

Smith, B. C. (1998). On knowing one's own language. (In Wright et al. (1998), 391-428)

Stich, S. (1983). *From folk psychology to cognitive science*. (Cambridge, MA: Bradford Books/MIT Press)

Wittgenstein, L. (1958). *The blue and brown books*. (New York: Harper & Row)

Wright, C. (1989). Wittgenstein's rule-following considerations and the central project of theoretical linguistics. (In A. George (Ed.), *Reflections on Chomsky* (pp. 233-64). Oxford: Blackwell)

_____ (1992). *Truth and objectivity*. (Cambridge, MA: Harvard University Press)

_____ (1995). Can there be a rationally compelling argument for anti-realism about ordinary ("folk") psychology?. *Philosophical Issues*, *6*, 197-221

_____ (1998). Self-knowledge: The Wittgensteinian legacy. (In Wright et al. (1998), 15-46)

_____ (2002). What could anti-realism about ordinary psychology possibly be?. *Philosophical Review*, *111*, 205-233

Wright, C., Smith, B. C. & Macdonald, C. (Eds.) (1998). *Knowing our own minds*. (Oxford: Clarendon Press)

Zimmerman, A. (2006) Basic self-knowledge: Answering Peacocke's criticisms of constitutivism. *Philosophical Studies*, *128*, 337-379