**Is "Self-Knowledge" An Empirical Problem? Renegotiating the Space of Philosophical Explanation**

Victoria McGeer

*The Journal of Philosophy*, Volume 93, Issue 10 (Oct., 1996), 483-515.

Stable URL:
http://links.jstor.org/sici?sici=0022-362X%28199610%2993%3A10%3C483%3AI%22AEPR%3E2.0.CO%3B2-X

THE JOURNAL OF PHILOSOPHY

VOLUME XCIII, NO. 10, OCTOBER 1996

# IS "SELF-KNOWLEDGE" AN EMPIRICAL PROBLEM? RENEGOTIATING THE SPACE OF PHILOSOPHICAL EXPLANATION*

Human beings are remarkable in their ability to perceive, discover, analyze, and understand things and, so, in their ability to come to know about each other and the world. But if philosophical tradition and common-sense intuition are to be believed, our success in knowing about things around us pales in comparison to our success in knowing ourselves. In certain ways, this is puzzling; for if coming to know things is an achievement, we seem to achieve best what we work at least. Self-knowledge, at least as it concerns our readily articulable beliefs, desires, feelings, thoughts, and other intentional states, seems peculiarly direct and certain. Reflecting on these states is sometimes necessary, of course, especially if we have been challenged in what we claim to believe, desire, hope, or fear; but, in the normal case, we do not worry ourselves about such things. If asked, we confidently say what is in or on our minds to others, and, as far as these matters are concerned, they confidently assume we know what we are talking about and that we have got it right. We are assumed to be in a privileged (epistemic) position with respect to the contents of our own minds; so others

* Versions of this paper have now been exposed to a number of helpful critics, and I have collected a number of debts in consequence. I would like to thank John Baressi, Richmond Campbell, Alison Gopnik, Martin Hahn, John Heil, Steven Horst, John Lachs, Duncan MacIntosh, Bob Martin, Bojana Mladenović, Denis Robinson, Eric Schwitzgebel, Paul Teller, and especially Randall Keen and Calvin Normore for detailed philosophical discussion and helpful editorial comments. I am grateful to the Killam Trust at Dalhousie University and the Social Sciences and Humanities Research Council of Canada whose postdoctoral support I enjoyed in seriatim while working on this project. I am also grateful to the Institute of Human Development at the University of California/Berkeley for providing a setting for ongoing interdisciplinary exchange and support.

0022-362X/96/9310/483–515 

grant us (as we grant ourselves) "first-person authority"—authority with respect to many of the everyday sorts of claims we make about ourselves.

This pretheoretic faith in authoritative self-knowledge, though widespread, has not gone unchallenged either in philosophy or in psychology. Most recently, a fairly technical debate in the philosophies of mind and language has given rise to a new series of questions about our putative self-knowledge, this time specifically with respect to the *content* of our intentional states. On the "externalist" side of this debate, philosophers have argued that the meanings of our words and the contents of our thoughts are determined in part by factors external to us.[1] Externalists differ as to what these factors are, but characteristically they include features of our physical environment (for example, underlying essences of stuffs with which we interact), our social environment (for example, shared conventions governing word use), or both. The problem is, whatever these environmental, content-determining features of our thoughts are, we could be ignorant of them (for example, of the underlying essence of water). And this seems to imply that there is nothing epistemically privileged about the first-person point of view: we might be quite confused about the contents of our own beliefs and other intentional states; perhaps we may even need to defer to experts in our linguistic community to help sort these confusions out. Thus, externalists seem faced with a dilemma not unlike one that faced an earlier generation of epistemological behaviorists:[2] either give up their favored account of mental contents or sacrifice the well-entrenched

[1] For seminal articles, see (1) Tyler Burge: "Individualism and the Mental," in P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy, Volume IV* (Minneapolis: Minnesota UP, 1979), pp. 73-121; "Other Bodies," in Andrew Woodfield, ed., *Thought and Object: Essays on Intentionality* (New York: Oxford, 1982), pp. 97-120; "Intellectual Norms and the Foundations of Mind," this JOURNAL, LXXXIII, 12 (December 1986): 697-720; (2) Donald Davidson: "The Myth of the Subjective," reprinted in M. Krausz, ed., *Relativism: Interpretation and Confrontation* (Notre Dame: University Press, 1989); "Meaning, Truth and Evidence," in R.B. Barrett and R.F. Gibson, eds., *Perspectives on Quine* (Cambridge: Blackwell, 1990), pp. 68-79; "Epistemology Externalized," *Dialectica,* XLV (1991): 191-202 (E. Rabossi, trans.); (3) Hilary Putnam, "The Meaning of 'Meaning'," *Philosophical Papers, Volume II: Mind, Language, and Reality* (New York: Cambridge, 1975), pp. 215-71.

[2] Such as Gilbert Ryle, *The Concept of Mind* (New York: Hutchinson, 1949). I follow Rockney Jacobsen ("Theory and Evidence in Self-knowledge," MS, Wilfrid Laurier University (1990)) in using the term 'epistemological behaviorism' to label "the doctrine that all the evidence for our intentional theories of persons, ourselves included, consists in publicly observable behavior" (p. 2). For further discussion of a modified version of this position, see Jacobsen, "Self-quotation and Self-knowledge," *Synthese* (forthcoming March 1997).

intuition that we have a peculiarly authoritative, though admittedly not infallible, knowledge of our own thoughts.

Some externalists have not been troubled by this choice. If the privilege accorded to first-person knowledge can be legitimized only by adopting an internalist, often called Cartesian picture of the mind and mental contents, then it is not in their view worth the price. Such externalists are happy, as was Gilbert Ryle, to "lose the bitters with the sweets of Solipsism" (*op. cit.*, p. 150).[3] Other externalists, whom I shall call *compatibilists*, have argued persuasively that first-person authority can be reconciled with an externalist account of meaning and belief once we reject an internalist conception of how claims to self-knowledge are justified. Hence, it is not simply a sweet of solipsism.

My sympathies lie generally with the compatibilists—philosophers such as Tyler Burge, Donald Davidson, John Heil, Crispin Wright, and Akeel Bilgrami, who claim that externalists are too quick to give up on first-person authority. They argue that the problem of explaining how we have privileged knowledge of our own intentional states is not, in essence, a problem about content. An agent may certainly be confused or wrong about something to do with the world: for instance, to cite Hilary Putnam's well-known twin-earth example, she may be confused about whether a glass of liquid in front of her is *water* or *twater*. But to suppose this translates into a confusion about her own mental states is to endorse—perhaps implicitly—those very features of the traditional picture of mind which externalists persuasively challenge. For it depends on construing beliefs and other intentional states as inner objects, albeit externally differentiated, which the agent herself can identify only by means of their subjectively accessible properties—most crudely, how they "appear" to the mind's eye. According to this picture, it is these objects (our representations of things) which we examine, or detect the properties of, to find out how things are in the world. Furthermore, it is these objects which we examine, or detect the properties of, to discover what it is in our own minds, thereby forming second-order beliefs about them which represent these facts about ourselves. But the picture fails, according to compatibilists, because there are no such objects before the mind. To find out about the world, to find out what we

[3] For externalists who either give up on or are pessimistic about the defensibility of authoritative self-knowledge, see, for example, Woodfield, *Thought and Object: Essays on Intentionality* (New York: Oxford, 1982), pp. vii-viii; Anthony L. Brueckner, "Brains in a Vat," this JOURNAL, LXXXIII, 3 (March 1986): 148-67; and Paul Boghossian, "Content and Self-knowledge," *Philosophical Topics*, XVII, 1 (Spring 1989): 5-26.

believe about the world, we turn our attention to outer objects and events, not inner ones. That is, we simply think about "the ordinary objects of the world that the outer eye registers and the heart loves."[4] Moreover, if the contents of our beliefs, desires, and other such states are puzzling to us, this is not because we do not know what to make of what is in our own minds; rather, we do not know what to make of surprising objects and events in the world (for example, whether the stuff in front of us is water—should be called 'water'—or not). We do not know how to integrate such things into our knowledge base.[5] This presents no threat to first-person authority, for the underlying problem in such cases is that we do not know (quite) what to believe; it is not that we believe something determinate, only we know not what.

What, then, would threaten the doctrine of privileged self-knowledge? There is, of course, a quite different sense in which an agent might be wrong about what she *really* believes—or desires or intends. This is the sense in which, as Wright says, an agent is "moved to opinions" about her own cognitive and emotional situation. Since her preoccupation here is not with understanding the world, but rather with knowing herself as a conscious, self-directed agent, the kind of content worries provoked by twin-earth scenarios are beside the point. As theorists, we should be concerned instead with accounting for an agent's ability to use and understand the conceptual repertoire of folk psychology, particularly with regard to interpreting her experiences and, so, explaining and justifying her own reactions to, and behavior in, the world. Here, questions about first-person authority are not only appropriate, but pressing; for there is no doubt an agent interprets her subjective experience of her own cognitive goings-on in the conceptual terms of folk psychology (beliefs, desires, hopes, intentions, and so on). But, absent the traditional picture of direct or privileged access, we have no satisfying account of why we should consider first-person knowledge to be so special—why, that is, in spite of all objections and provisos, it still seems to us peculiarly direct, and why agents' judgments should be considered peculiarly authoritative.[6]

[4] Davidson, "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association*, XXX (1986): 441-58, here p. 453.

[5] I make this argument in much greater detail in "On Living Languages and Dead Thought-experiments: What Twin Earth Really Tells Us about the Meaning of 'Meaning'," MS, Vanderbilt University (1995).

[6] These two theses need not be—in fact, I shall argue, cannot be—connected in the way tradition supposes. There may be interesting connections nonetheless. I make this case in section II below.

Some philosophers, such as Wright and Bilgrami, have hoped to make headway on this problem by abjuring epistemological accounts of self-knowledge altogether.[7] Within the epistemological tradition, our faith in first-person authority is held to be justified (or not) according to the justificatory relationship argued to obtain between our first-order intentional states and the second-order beliefs we form about them. The Wright-Bilgrami alternative is to recognize how our faith in first-person authority itself underlies and shapes the structure of our day-to-day interactions; so whatever justification it merits derives from the success (Bilgrami says human indispensability) of these interactions. The philosopher's project is thus redefined. It is to defend first-person authority not internally, so to speak, but rather externally, by showing how our ways of interacting with one another entitles us to assume that, normally speaking, our judgments about ourselves are systematically and significantly reliable. This obviates the need for a *philosophical* analysis of how we come to make such judgments. As Wright explains:

> Since the telos...of the practice of ascribing intentional states to oneself and others is mutual understanding, the success of a language game that worked this way would depend on certain deep contingencies. It would depend...on the contingency that taking the self-conceptions of others seriously, in the sense involved in crediting their beliefs about their intentional states, as expressed in their avowals, with authority, will almost always tend to result in an overall picture of their psychology which is more illuminating—as it happens, enormously more illuminating—than anything which might be gleaned by respecting all the data except the subject's self-testimony. And that in turn rests on the contingency that we are, each of us, ceaselessly but—on the proposed conception—subcognitively moved to opinions concerning our own intentional states which will indeed give good service to others in their attempt to understand us. Thus, we do not cognitively interact with states of affairs that confer truth upon our opinions concerning our own intentional states; rather, we are inundated, day by day, with opinions for which truth is the default position, as it were. They count as true provided that we hold them and that no good purpose is served, in another's quest to find us intelligible, by rejecting them (*op. cit.*, pp. 632-33).

Bilgrami's view is more encompassing still. While he agrees it must be our default position to assume we have first-person authority, he

[7] Wright, "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy, and Intention," this JOURNAL, LXXXVI, 11 (November 1989): 622-34; Bilgrami, "Self-knowledge and Resentment," MS, Columbia University (1994) and *Belief and Meaning* (Cambridge: Blackwell, 1992).

argues that this is not simply a precondition of playing the language game of intentional ascription successfully. Rather, it is a stance we must adopt if we are to treat one another as moral agents. After all, we can only be held responsible for our acts if we know what we are doing—that is, if we know what acts we are performing. But in order to know what acts we are performing, we must know our own intentional states; for it is the fact of these states *causing and being reason* for our acts which makes them into (identifies them as) the *intentional* acts that they are. Consequently, if we sacrifice the intuition that, generally speaking, we know our own intentional states, we sacrifice our right to treat one another as moral agents, cognizant of and responsible for what we do.[8]

In my view, there is something genuinely attractive about accounts like these which shift the problem away from traditional epistemology into what Bilgrami has (re-)called *philosophical anthropology*. It allows philosophers to concentrate on what is at stake in our being accorded the respect and responsibility of self-knowers, while, at the same time, leaving the exploration of those cognitive and subcognitive processes which realize this competence to the specialized sciences. This is *not* to advocate a separate peace, however. As Bilgrami suggests, certain ideas about the kinds of cognitive processes characterizing self-knowers may seem more or less attractive once philosophers give full-flight to considerations relevant to answering their questions about meaning, act, intention, agency, responsibility, and freedom. Indeed, he would like to claim that psychological models that presuppose any kind of mechanism for self-knowledge (whether conscious or unconscious) that puts us in the passive role of being mere recipients of information about our own first-order intentional states miss something profound about the connection between our capacity for self-directed thought and action (that is, our agency) and self-knowledge.

Once again, I am deeply sympathetic to these reflections, as I hope will become evident from the account of self-knowledge I develop in section II. Before heading in this direction, however, there is some critical ground that first needs to be covered. For the move I want to make in this discussion, the move I believe Bilgrami is making, may seem misdirected or, worse, a fundamentally irrelevant philosopher's shuffle if the shape of the terrain is not well understood.

For instance, a prima facie reasonable criticism of Bilgrami's approach would be the following: suppose it is true that our capacity

[8] Bilgrami, "Self-knowledge and Resentment," p. 6.

for self-knowledge is bound up with our capacity for responsible agency in the way Bilgrami suggests. Does the fact that we now react to one another *as if* we had the latter capacity (and so the former) tell us anything about what capacities we actually *do* have? Surely this is just the sort of presumption about our psychological make-up that empirical investigation could lead us to question. In fact, as Bilgrami himself notes, in quarters where certain medical models of psychological functioning dominate, such questioning has already occurred (*ibid.*, p. 13). Thus, the empirically minded could simply agree with Bilgrami's observation that our personal and societal values are deeply connected to our acting and reacting as (self-knowing) agents. But this just means the consequences will be far-reaching if the psychological facts about us turn out to contradict our intuitive (and socially buttressed) convictions. That is, we may be forced to revise radically our self-conceptions, retool our normative modes of interaction (including our language games of intentional ascription), give up, in fact, on certain important tenets of folk psychology. But in the empiricists' view, the magnitude of this task, though daunting, is no argument for our having authoritative self-knowledge. In fact, these critics might say, all the more reason for our exploring these matters empirically: the better we understand how and to what extent we are reliable self-knowers (hence, properly in control of our actions and reactions), the better able we shall be to interact with one another, both on a personal level and, even more importantly, on a societal level in terms of the institutions we develop for education and reform.

Now, this objection presupposes that there is an appropriate division of labor between philosophers and scientists which has something like the following character: philosophers, like Wright and Bilgrami, operate at a conceptual level tracing the interconnections between how we think about agency and self-knowledge and the practices that structure our interactions. Scientists, on the other hand, find out the facts about us and thereby determine whether our thoughts and practices are based on reality or on mere prejudice. Sensible as this division may sound, however, it seriously mischaracterizes the contribution philosophers and scientists can make, and have made, to each other's research efforts. I say this not to insulate the problem of self-knowledge from scientific exploration, as if it really were just a philosopher's problem after all, admitting of an a priori solution. I say this, rather, because the proposed division of labor assumes a naive empiricism which has been rightly and forcefully criticized. Scientists may well be presented as "constructing the-

ories on the basis of evidence" or "putting them to the test"; but, if so, it must be acknowledged that these activities are far more complicated than such simple characterizations suggest. Theories evolve in relation to a number of interests and constraints, partly internal and partly external, partly conceptual and partly evidential. Moreover, how we view "the evidence" depends very much on our conceptual resources, and these, of course, are provided in part by the theories themselves. This is not to advocate antirealism. It is merely to say that there is no easy separation between the two levels of inquiry—conceptual and empirical; on the contrary, there is dynamic interplay between myriad kinds of considerations and room always, then, for philosophical contribution.

Nowhere has the fruitfulness of this exchange been more evident than in the relation between philosophers' and psychologists' work on mind and behavior. For instance, inspired by much recent work in the philosophy of mind, many psychologists have begun to (re)organize their research programs around the philosophical view that our common-sense attribution of intentional states is guided by our "folk psychology," a "protoscientific theory" developed and used to explain and predict our own and each others' behavior.[9] In keeping with their nature, psychologists have dubbed this the "theory theory,"[10] indicating their view that this idea of us ordinary folk using such a theory in our day to day interactions is itself a theory open to empirical investigation by psychologists. This theory theory has many interesting features. With regard to the problem of self-knowledge, the hypothesis is that we use our folk notions of the underlying causes of behavior to attribute mental states and processes to ourselves just as much as we use them to make attributions to others. In other words, our self-judgments are as mediated by theory (folk theory) as our third-person attributions. Moreover, because there is no unmediated access to the contents of our own minds, there is no good reason to endorse a traditional epistemological account of privileged self-knowledge.

This hypothesis about self-knowledge has not gone "untested," and the body of psychological evidence adduced to support it is fairly im-

[9] Paul Churchland, *Matter and Consciousness* (Cambridge: MIT, 1984); Daniel Dennett, *The Intentional Stance* (Cambridge: MIT, 1987); Jerry Fodor, *Psychosemantics* (Cambridge: MIT, 1987), *A Theory of Content and Other Essays* (Cambridge: MIT, 1990).

[10] Louis Moses and H.J. Chandler, "A Traveler's Guide to Children's Theories of Mind," *Psychological Inquiry* (in press); A. Gopnik and H.M. Wellman, "Why the Child's Theory of Mind Really Is a Theory," *Mind and Language*, VII (1992): 145-71; J. Perner, *Understanding the Representational Mind* (Cambridge: MIT, 1991); H.M. Wellman, *A Child's Theory of Mind* (Cambridge: MIT, 1990).

pressive. In section I, I shall review some of this research to emphasize in particular how deeply it seems to cut against our common-sense intuitions about first-person authority. From this it is clear that we need a better account of self-knowledge than is provided either by our well-entrenched folk prejudices or by traditional philosophical accounts that may well have inspired them (or been inspired by them). For what this research seems to be showing is precisely what tradition counterpredicts, namely, that there are similar patterns of error across first- and third-person attributions. In other words, not only do we go wrong about ourselves; but when we go wrong, we tend to do so in exactly the same ways as we go wrong about others.

Be the psychological evidence as it may, I think the current state of inquiry calls for continued interplay between psychology and philosophy; for as things now stand in the two disciplines, we are facing something of a puzzle. On the one hand, in rejecting traditional accounts of self-knowledge, philosophers like Wright and Bilgrami have given us very persuasive arguments for subscribing to a cleaned-up doctrine of first-person authority. In their version, this is the thesis that we adopt a "positive presumptive" attitude toward first-person claims about intentional states. That is, we assume such claims are true, doubting any particular claim only if we have very good reason to do so. Psychologists, on the other hand, seem to reject tradition by telling us that, according to their data, this positive presumptive attitude is illegitimate. That is, whatever attitude we adopt toward third-person attributions is the attitude we ought to take toward first-person ones, since they are liable to the same kinds of error. Whom are we to believe?

The conservative response would be to assume that this disagreement is more apparent than real. After all, the philosophers are not claiming that first-person judgments should be considered infallible, and the psychologists are not saying that we have *no* reason to favor first-person claims. Perhaps, then, their views can be reconciled. It may simply be a matter of calibrating their discussions in terms of the kinds of mental states and processes under consideration, or the kinds of situations in which first-person judgments are made, or the kind of self-judging subject in question (that is, child or adult). Yet, important as it is to sort all this out, I think a more fundamental disagreement can still be discerned.

In section II, I shall sketch the shape of this disagreement and suggest that a more dramatic resolution is required. Throwing my lot in with the philosophers, I shall argue that their conceptual work on self-knowledge opens up an entirely new way of thinking about many

of our first-person claims. But while this new way of thinking is inspired by philosophical considerations, it is driven by psychologists' empirical findings. For in my view, our project must be to develop an account of self-knowledge that can subsume this psychological work on first-person error while leaving the inalienable, intentional authority of agents intact. In the end, I believe we can achieve this goal, but only by significantly revising our theoretical understanding of the practice of folk psychology, particularly with regard to its role in our cognitive and social development. This conclusion, with its suggestion for new directions in research, may seem fairly radical, departing as it does from the received view of folk psychology in both philosophy and (academic) psychology; but I hope the advantages it may offer in allowing us to knit together a number of different concerns will recommend its serious consideration and further exploration.

## I. COMING TO GRIPS WITH FIRST-PERSON ERROR: FROM ORDINARY INTUITIONS TO SCIENTIFICALLY GROUNDED CONVICTIONS

In our everyday dealings with one another and the world, we seem pretty firmly committed to two conflicting intuitions. On the one hand, we have an abiding faith in first-person authority: knowledge of our own states is peculiarly direct and certain; on the other, we do recognize that we are fallible creatures: we may get things right much of the time (or right enough to survive), but nothing we do is entirely free from error, even if that something is making judgments about our own mental goings-on. This conflict of intuitions makes it hard to get a grip on the nature and scope of first-person knowledge. For instance, while we do expect occasional instances of first-person error, these seem particularly hard to chart. This is not just because the sort of evidence we find for them in everyday encounters is rather vague, consisting of apparent dissociations between how a person behaves and what he has claimed about his own beliefs, desires, feelings, and so on. In addition, our strongly held, commonsense conviction that individuals usually get their mental states right fights, in any particular case, against believing an agent has erred. Thus, even if such dissociations become apparent to the person himself, causing him to reflect on how and what he thinks about his own states of mind, we are inclined to give both others and ourselves plenty of latitude in elaborating explanations of our behavior that leaves first-person claims about underlying mental states more or less intact.[11]

[11] In fact, I think this generosity is not unwarranted; but not for the straightforward reason that agents are generally right (in the standard epistemic sense) about their first-order states. I return to this crucial idea in section II below.

Still, the dissociations between what individuals say and how they act are persistent and intriguing enough to warrant systematic investigation; and this, in turn, has prompted psychologists to question seriously our common-sense faith in first-person authority. Sigmund Freud, of course, was famous for his theories about unconscious motivation. More recently, R. Nisbett and T. D. Wilson[12] have catalogued a variety of experiments that demonstrate repression may not be the only, or most common, explanation for agents' mistaking their own understanding of why they react and act as they do. In fact, in general, according to Nisbett and Wilson, people have no reliable subjective access to the cognitive processes underlying their judgments and their behavior. They can easily be misled about what factors influence their evaluations, their choices, their behavior, even though they will happily volunteer explanations of these as if based on knowing acquaintance with the internal processes that led to them. Similar reports in the study of emotion demonstrate that people will claim they directly experience affective states that are "appropriate" in various contexts (fear, for instance, or sexual excitement), even though the underlying physiological state responsible for these experiences in all cases is nothing more than a mild, stimulant-induced discomfort.[13]

In developmental psychology, Hanz Wimmer, Josef Perner, Alison Gopnik, Henry Wellman, and others have made some interesting discoveries relating directly to our first-person experience of intentional states. In a series of experiments with young children, they have found that three-year olds (unlike four-year olds) seem unable to make correct judgments about their own beliefs once they discover them to be false. In one experiment, for instance, children are shown a closed, but easily recognizable candy box and asked by the experimenter what they think is inside. The children all say "candy." When the jury-rigged box is opened to reveal pencils, the children are surprised. The box is closed up again and the experimenter asks: "When you first saw this box all closed up like this what did you think was inside?" Three-year olds consistently respond: "pencils."[14]

[12] "Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, LXXXIV, 3 (May 1977): 231-59. See also Nisbett and L. Ross, *Human Inference: Strategies and Shortcomings of Social Judgement* (Englewood Cliffs, NJ: Prentice Hall, 1970).

[13] S. Schachter and S. Singer, "Cognitive, Social and Psychological Determinants of Emotional State," *Psychological Review*, LXIX (1962): 379-99—as reported in Alison Gopnik, "Psychopsychology," *Consciousness and Cognition*, II (1993): 264-80.

[14] Gopnik and J.W. Astington, "Children's Understanding of Representational Change and Its Relation to the Understanding of False Belief and the Appearance-Reality Distinction," *Child Development*, LIX (1988): 1366-71.

By contrast, these children have no difficulty making correct judgments about their previously held *true* beliefs, even if these are not in fact the beliefs they currently hold. In a closely related experiment, children were shown a candy box that was opened (after questioning) to reveal candy. In front of the children, the candy was removed and replaced with pencils; then the box was closed up again. The children were asked the same question as in the previous experiment: "When you first saw the box all closed up like this, what did you think was inside it?" Although the children knew the box was currently full of pencils, they still gave the correct answer: "candy."[15] Thus, in cases where children are unable to lay claim to a previously held false belief, their mistakes do not appear to stem from a simple limitation on memory.

There are other experiments that reveal unexpected difficulties in children's awareness of their own first-order states—unexpected, partly because of our adult intuitions concerning the ease and reliability of first-person knowledge of our minds and partly because children of this age are surprisingly good at a number of other cognitive tasks that seem equally difficult. Thus, for instance, though able to engage in pretend play (for example, where they treat a banana as a telephone)[16] or remember events from up to a week in the past,[17] Gopnik and Virginia Slaughter have shown that thirty- to thirty-six-month olds misreport their previously held (and expressed) desires as soon as those desires have been realized. That is, these children will simply deny having had their previously held desires.[18] The problem here seems closely to mirror the problem in reporting previously held false beliefs, even though the capacity for understanding the nature of desires and intentions seems to develop before the capacity for understanding the nature of beliefs and other complex cognitive states. (Three-year olds do not have the same difficulties with the "desire task" that two-and-a-half-year olds do.) Nevertheless, in all these cases, as Gopnik points out, for adults, "the phenomenological feel of privileged access is particularly strong" (*op. cit.*, p. 272). She adds:

[15] H. Wimmer and M. Hartl, "Against the Cartesian View on Mind: Young Children's Difficulty with Own False Beliefs," *British Journal of Developmental Psychology*, IX (1991): 125-28.

[16] See Alan Leslie, "Some Implications of Pretense for Mechanisms underlying the Child's Theory of Mind," in J.W. Astington, P. Harris, and D. Olson, eds., *Developing Theories of Mind* (New York: Cambridge, 1988), pp. 19-46.

[17] Katherine Nelson, *Event Knowledge, Structure, and Function in Development* (Hillsdale, NJ: Erlbaum, 1986).

[18] "Children's Understanding of Changes in Their Mental States," *Child Development*, LXII (1991): 98-110.

> If you asked me how I know that I believed that there were candies in the box 30 seconds ago, I will reply with some sort of introspective account; I looked in my mind and there was the belief. And yet children are reporting a very different belief, in fact, a belief they could not possibly have had, with equal authority and certainty.
>
> Where do the children's mistaken beliefs come from (*op. cit.*, p. 272)?

One suggestive piece of this puzzle comes from the discovery that the same pattern of error in self-reports occurs in young children's claims about the intentional states of others.[19] Similarly, in the studies cited by Nisbett and Wilson, it seems that the kinds and frequency of errors adults make in describing their own cognitive processes are ones they also make in evaluating what will affect the judgment and behavior of others (*op. cit.*, pp. 247-49). Furthermore, according to Nisbett and Wilson, "the evidence suggests that people's erroneous reports about their cognitive processes are not capricious or haphazard, but instead are regular and systematic" (*op. cit.*, p. 247). The same can be said about children's errors.

These two things taken together—the symmetry of first- and third-person error coupled with its systematicity—suggest to psychologists working in these fields that, in both the children's case and the adults', their judgments about mental states and processes are affected by their background beliefs about how these are caused and how they in turn cause behavior. This has led Nisbett and Wilson to propose that people only engage in something that "feels like introspection." In fact, they argue,

> ...when people are asked to report how a particular stimulus influences a particular response, they do so not by consulting a memory of the mediating process, but by applying or generating causal theories about the effects of that type of stimulus on that type of response. They simply make judgements, in other words, about how plausible it is that stimulus would have influenced the response (*op. cit.*, p. 248).

[19] In another series of experiments, the so-called "false-belief task," children are required to recognize that someone else's belief will be false and that her actions will be based on that false belief. In a paradigmatic case, the agent in question (a puppet named "Maxi") puts a candy in box *A* and leaves the room. While the children are watching, the candy is moved from box *A* to box *B*. They are then asked: "Where will Maxi think the candy is?" "Where will Maxi look for the candy when he comes back?" Again, three-year olds consistently say both that Maxi will think the candy is in box *B* (where it really is) and that he will look for it there—Wimmer and Perner, "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception," *Cognition*, XIII (1983): 103-08. For further discussion of this parallel between first- and third-person ascription, see Gopnik.

Of course, since adults apply the same causal theories to themselves and others, one would expect to see the same kinds of errors in both first- and third-person claims, reflecting particular problems with the theories themselves. Similarly in the children's case, Gopnik argues that their errors reflect their "mistaken implicit theory of how the mind works" (*op. cit.*, p. 272). Three-year olds, for instance, seem to have what has been aptly termed a "copy theory" of belief (cf. Wellman). According to the copy theory, states of the world (for example, where objects are located) are impressed directly on the mind. This leaves no room for interpretation and, hence, for misrepresentation. If someone (including oneself) has a belief about what is in the box, in the three-year old's view, it must be a copy of what is *really* in the box.[20]

While these proposals seem to fit the data of agents' errors in self-reports (as well as their errors in claims about others' mental states and processes), more needs to be said to account for the phenomenology of first-person experience and to assess what, if any, special authority these experiences might yet confer on first-person beliefs and claims about the mental. Nisbett and Wilson present a rather sketchy and divided picture of how we know our own minds. On the one hand, as I mentioned before, they argue that the feeling of direct or subjective access to the workings of our minds (that is, to our actual cognitive processes) is illusory; on the other, they claim we fall prey to this illusion because we do have access to a "storehouse of private knowledge"—knowledge that consists in awareness of relatively momentary things, like focus of attention and current bodily sensations, and relatively lasting things, like emotions, intentions, evaluations, plans, character traits, and so on (*op. cit.*, pp. 255-56). Knowing about these things improves our ability to utilize theories and generalizations in gauging the causal impact of various stimuli on our behavioral responses. But lest we be tempted to find much support in this for the doctrine of first-person authority, Nisbett and Wilson caution that such "private access to content" may work against us as well: "Occasionally, noninfluential stimuli may be more vivid and available to the individual than to an outside observer... and thus the observer might sometimes be more accurate by virtue of disregarding such salient but non-influential stimuli" (*op. cit.*, p. 256).

Be that as it may, Nisbett and Wilson may still be painting too rosy a picture of first-person access. They admit, for instance, that some

[20] Wellman, chapter 9; Gopnik, "How We Know Our Own Minds: The Illusion of First-Person Knowledge of Intentionality," *Behavioral and Brain Sciences*, XVI (1993): 1-14.

"mystery" attaches to "why people are so poor at telling the difference between private facts that can be known with near certainty and mental processes to which there may be no access at all" (*op. cit.*, p. 255). In fact, what may seem more mysterious is why we should think people have straightforward access to their mental states if not to their mental processes, particularly in light of the fact that in *both* cases subjects seem often to feel as if they do. Here, the developmental evidence becomes important because it shows that this distinction between access to processes and access to states is not so clear-cut. Indeed, it seems that in both cases our experience of our own minds is mediated by our theoretical preconceptions. The only difference between young children and adults is that the adult folk theory does a better job than the three-year old's theory at mapping the causal structure of the underlying psychological states and processes that mediate between sensory inputs and behavioral outputs. The theory is better, but not of course perfect, since there are, as Gopnik points out, "areas of our mental life, like motivation, that [adult] folk-psychology does not understand very well."[21]

The developmental evidence leads Gopnik to propose a more plausible picture of how we come to have the experiences we do of both mental states and processes than the bipartite scheme of access/no access that Nisbett and Wilson propose. Modeling her account on an information-processing view of perception, Gopnik suggests our "introspective" experiences of our own mental states, like our perceptual experiences of physical objects, are the result both of (relatively insulated) bottom-up transformations of information received as proximal stimuli and of top-down effects generated by "expectations, preconceptions, and higher-order beliefs"—in effect, our theories of how things work in the relevant domain, whether it be world or mind (*ibid.*, pp. 273). So just as it feels phenomenologically as if we directly *see* a red apple before us or, as in the case of a chess expert, a developing threat to the queen, so it feels as if we directly experience our belief that there is candy in the box or our desire that Aunt Tillie's umbrella be hidden from Uncle Frank. Phenomenological directness, in other words, is no indicator of direct or unmediated access, even though it certainly gives us this illusion.[22]

[21] "Psychopsychology," p. 273.

[22] Philosophers will no doubt recognize in this a thesis similar to Paul Churchland's concerning the plasticity of perception and the mutability, therefore, of our experience of our own minds: "Self-perception consists in the disposition-governed occurrence of conceptual responses to one's internal states, responses made within

Those searching to legitimize the intuition of first-person authority may thus take no comfort from the role more sophisticated psychological theories accord to our purportedly direct phenomenological experience of our own minds. We may have "privileged access" in the sense that the underlying psychological states mediating our intelligent responses to a changing world (our beliefs, desires and other first-order states) also produce in us psychological experiences that give us some awareness of these states; but given how these experiences seem to be shaped by our folk theories of how the mind works (what mental states are, what kinds of circumstances produce them, what behavioral consequences they have, and so on), our knowledge of the underlying states that produce them will be circumscribed by the adequacy of our theories. We thus have a nice account of first-person error, but only, it seems, at the cost of significantly reducing, if not eliminating, the basis for our common-sense faith in first-person authority.

There is, of course, the usual consolation to be found in undermining the "privileged access" account of authoritative self-knowledge. For if first-person experience gives us no particular grounds for claiming a peculiarly authoritative knowledge of our own states of mind, not having such experience of the minds of others gives us no particular grounds for thinking our judgments about their states must be relatively uncertain. Indeed, according to Gopnik, this is precisely the result we want. For we need to explain not just the occurrence of error in the first-person case, but also why there is a particular *symmetry* in patterns of error across first-person claims and third-person attributions—again, not just in the children's case, but in the adults' as well. Furthermore, Gopnik claims (provocatively, but not without merit), the relative uncertainty we are trained to attach to third-person attributions may be justified by the Cartesian

---

whatever matrix of self-understanding one has developed or acquired.... It is therefore possible that such responses might come to be made within the framework of a more penetrating theory of our internal states and activities. And...the difference between the old inner vision and the new may be substantial"—*Scientific Realism and the Plasticity of Mind* (New York: Cambridge, 1979), pp. 116-18. Churchland, of course, uses this to buttress his attack on the presumed invulnerability of folk-psychological concepts such as beliefs, desires, and other intentional states to an advancing scientific understanding of cognitive processes. It is worth noting that while Churchland and Gopnik disagree about the utility of folk-psychological concepts and generalizations in delimiting the underlying causal structure of our psychological states and processes, they at least agree that this *is* the business of folk psychology—to represent adequately our "internal reality"—Churchland, pp. 95-100; Gopnik, "Psychopsychology," pp. 272-73, and "How We Know Our Minds...," p. 12.

picture of our distant and inferential reach to other minds, but this distance itself does not seem to be phenomenologically justified. She writes:

> ...purely phenomenologically, our perception of mental states in others is, at least much of the time, no less immediate than our perception of our own mental states. I "see" my son's hunger or my friend's disappointment just as directly as I see my own. Indeed, if we imagine what a purely physical perception of other people would be like, a perception from which we then inferred their mental states, it seems as bizarre as imagining ordinary visual perception as an inference from an uninterpreted pattern of light and dark. Imagine seeing the other people around you at the dinner table, say, as bags of skin stuffed into bags of cloth, with two small restless black spots near the top and a hole underneath that emits noises. This is a mad view. At the most immediate phenomenological level, particularly with familiars, there is no other minds problem. Like so many other problems, it only emerges when we start to think.[23]

When we start to think as folk philosophers of mind, we develop, according to Gopnik, a naive theory of how knowledge of first-order mental states is acquired—a theory that both builds on the undeniable fact that our mental states are stashed away somewhere inside our heads and, crucially, is *perceptually* based. It thus seems plausible to us that, despite how easily we seem to "read" other people's states of mind from their expressions and so on, such states can really only be transparent to those who have them since only they have an unimpeded view of them (their heads do not get in the way). And so we arrive at the traditional asymmetry between first- and third-person knowledge of mental states, an asymmetry which has obvious implications for the degree of authority we think appropriate to invest in first- and third-person claims.

In Gopnik's view (and in the view of others), evidence from psychological investigations demands that this naive theory be replaced. Our knowledge of our own mental states is as mediated as our knowledge of others'—mediated, that is, by our background folk-psychological theory of the causes and consequences of intelligent behavior. Of course, the causal routes through which we garner information about our own versus other minds cannot be the same, just as the means by which we (normally) garner information about the expression on our own faces (proprioception) is not the same as the means by which we garner information about the expression on

[23] "Psychopsychology," p. 269.

another's face (visual perception). And while this difference means we *experience* our mental states as our own, and others' states as theirs (so not in the same way), we must not forget that our experiences in both cases are shaped by the same interpretive schema.

What implications has this for our common-sense faith in first-person authority? Perhaps this intuition can be salvaged if we have reason to think experiences generated by information reaching us through the inside passage, as it were, are more complete and, hence, better suited to our making correct judgments about our own beliefs, desires, intentions, and the rest. Perhaps, in other words, we have the kind of residual authority Ryle was inclined to invest in our first-person judgments simply because they are based on more of the same kind of information available to others (even if the means by which we get it, pace Ryle, is different). After all, it is not all that easy to get away from ourselves. But, of course, more of the same kind of information is not always a good thing if the information is being filtered through an interpretive scheme that is, in some way, distorting (as in the case of three-year olds). Yet this, too, may be turned to theoretical advantage; for this gives us a plausible way to account, not just for variations in degree of first-person competence between young children and adults, but even for variations (normally less extreme) among adults themselves. We, thus, can relax the idealizing assumption that all minds are equal in first-person knowledge.

All in all, then, we seem to have a pretty good account of self-knowledge before us: we can allow for first-person error; we can explain the symmetry between first- and third-person patterns of error; we can account for differences in first-person competence; we preserve the phenomenology of first-person subjective experience; we allow for the phenomenology of directly detecting others' intentional states (by denying that all we "see" in the case of others is how they behave, our judgments of their mental states depending on inferences we make, either from analogy or to the best explanation). And we get all this simply by jettisoning an understandable, though naive, picture of the mind and mental contents. It is this picture, moreover, that seems to explain the temptation we feel to characterize judgments about our own states of mind as peculiarly authoritative, arrived at in a specially direct way and therefore challengeable only in special circumstances. Give up this picture and we are free to think of the asymmetry between first- and third-person access to mental states in scientifically acceptable ways, while yet remaining skeptical of the epistemological mileage philosophers have traditionally thought to make out of it.

## II. PHILOSOPHY (AND FIRST-PERSON AUTHORITY) REDUX

We have seen how the traditional picture of mind—the picture according to which "the mind is a theatre in which the conscious self watches a passing show...of appearances, sense data, qualia"[24] and where first-person authority stems from having the only seat in the house—has come under attack from two different and apparently competing directions.

First, the wave of criticism inspired by externalist accounts of content has made us suspect that thinking about self-knowledge as knowledge of some*thing* (inner states) gained by some species of observation is radically misconceived. That is, it may be one thing to view self-knowledge as dependent on having by and large true beliefs about one's own intentional states; it is quite another to view this condition as the result of reliably identifying one's first-order states by perceiving in them their individuative conditions. For as compatibilists argue, this concedes too much to the idea that knowledge of external things is mediated by representations, understood as subjectively accessible objects before the mind which *re-present* reality much as pictures do. Banish this "myth of the mental" (as Davidson calls it) from our philosophical repertoire and we can accept in full generality the consequences of adopting an externalist picture of the mind. If first-order intentional states are determined to be what they are in part by factors external to individual agents, then so will be second-order beliefs about those intentional states. There is simply no gap for a skeptic to exploit between the agent's first-order states and her second-order beliefs about them. As Davidson[25] says, "whatever is responsible for the contents of our thoughts, whether known or not, is also responsible for the content of the thought that we have the thought" (*ibid.*, p. 1). Heil[26] elaborates:

> We must understand theories of content as setting out conditions that agents must satisfy if they are to have contentful states of mind. Their satisfying these conditions need not be a matter of their recognizing them to be satisfied.... Beliefs about the contents of one's own thoughts, then, need not be based on beliefs about whatever it is that fixes the contents of those thoughts. The contents of second-order thoughts are fixed, just as are the contents of first-order thoughts, by the obtaining of appropriate conditions (*ibid.*, pp. 250-51).[27]

[24] Davidson, "Knowing One's Own Mind," p. 453.

[25] "Reply to Tyler Burge," MS, University of California/Berkeley (1988). Note: this MS is a longer version of the reply published in this JOURNAL, LXXXV, 11 (November 1988): 664-66.

[26] "Privileged Access," *Mind*, XCVII (1988): 238-51.

[27] See also: Burge, "Individualism and Self-knowledge," this JOURNAL, LXXXV, 11 (November 1988): 649-63; Davidson, "Knowing One's Own Mind"; Bilgrami, *Belief and Meaning*.

This style of response may sound reasonable as far as it goes. It is important, however, to separate the compatibilists' critical points against a more traditional account of self-knowledge from the positive alternative that I have quickly sketched here. For in my view, invoking a structural hierarchy of intentional states in this context is both confusing and (ultimately) self-defeating. My reasons are somewhat complex.

In the first place, this analysis makes it seem as though being consciously aware of, and even expressing a first-order state—for example, there is water in the glass—depends on having formed (perhaps unconsciously) a second-order belief about that first-order state. But is this right? It seems a fundamental insight of the "externalist turn" that thinking (and talking) about the world consciously does not consist in an unconscious, second-order process of thinking about thoughts about the world. Rather, it is just a conscious first-order process of thinking (and talking) about the world.

In the second place, the implicit assumption seemingly buried in this, that normally speaking we have first-person authority because normally speaking there is a relatively direct and unproblematic connection between first- and second-order thoughts, is as bad a structuring premise for externalist accounts of first-person authority as it was for more traditional views. For whether this connection is made out in old-fashioned epistemological terms (by invoking the mind's eye) or in new-fangled causal ones (by invoking a "reliable" internal tracking mechanism), we can only presume we have first-person authority to the extent we can assume this connection is direct and trouble free. Now, in so far as we acknowledge the phenomenon of first-person error, we have acknowledged that this connection is not entirely trouble free. But the real threat to traditional first-person authority comes from psychological investigations that reveal the *systematicity* of such errors, as well as their pattern of serial extinction in the course of development. For this forces us to question the *directness* of this putative connection between first-order states and second-order beliefs about them, especially given the plausible hypothesis that how we experience our own psychological states is significantly affected by the background folk-psychological theory we develop over time to predict and explain the causes and consequences of intelligent behavior. Thus, the traditional picture of mind comes under attack from a second direction. As Gopnik writes,

> ...real psychopsychology is more like real psychophysics and perception than like the folk or philosophical picture of privileged access. We can be and often are mistaken about our own mental states, in the sense

> that our subjective experience does not match the structure of our underlying psychological states. And often these illusions are the result of top-down influences on subjective experience. Our theories of the world or the mind, our general knowledge, our other beliefs and expectations, our expertise, play as great a role in shaping our subjective experiences as the mental states or physical states those experiences are apparently about.[28]

It is interesting to note the dialectic that emerges from these two criticisms of the traditional picture of mind. Given the doctrine of externally determined contents, and given that many externalists assume authoritative self-knowledge hinges on having by and large true second-order beliefs about our own intentional states, they address the problem of the heritability of first-order content by arguing that we do not form our second-order states by (somehow) judging the content of our first-order states (as if by seeing). But how else to explain this connection? The proposal (made explicitly by Wright and perhaps implicitly by Davidson) is that, as a matter of brute fact, we are, according to Wright, "ceaselessly but subcognitively moved" to "opinions [about our first-order states] for which truth is the default position..." (*op. cit.*, p. 633). Yet, the growing body of psychological evidence discussed in section I indicates that, in important and predictable ways, truth should *not* be the default position. Consequently, we are forced to reconsider the plausibility of Wright's epistemically unmediated link. Otherwise, it is hard to explain how the opinions to which our psychological experiences give rise are so clearly affected by our developing (and in some significant respects, inaccurate) theoretical views of how minds work. In other words, the traditionalists may have been right after all to analyze self-knowledge within an explicitly epistemological framework, but wrong to suppose the "observations" we make of our own psychological states are not theory laden.

Is there no further move to be made in this dialectic? Is a philosopher like Wright simply ignorant of pertinent psychological findings and so naive in his views? As I mentioned earlier, I believe this judgment would miss the power and richness of conceptual analyses that attempt to integrate diverse kinds of considerations in order to provide a picture of ourselves that is satisfying in a number of different dimensions. Wright, for instance, may have his cognitive mechanisms wrong, but he is surely right in his observation that we have a well-entrenched folk-psychological practice (or "language game") of

[28] "Psychopsychology," p. 273.

intentional attribution for which there are certain preconditions of success. One precondition he identifies is that people are able to rely on our utterances to guide them in making attributions that will explain and predict our behavior (*op. cit.*, p. 632). If this were generally not the case, the practice would surely have died out.

This conclusion may seem hyperbolic. For if our behavior did not *generally* mesh with the claims we make, people would simply have no reason to trust what we say or (if they think us sincere) credit us with first-person authority. Our first-person claims would need to be discounted in their attempts to make sense of us from what Daniel Dennett[29] calls *the intentional stance.* But is this imagined strategy of systematically ignoring our first-person claims in order to make sense of us via intentional ascription really credible? There are reasons to doubt it. The success of folk psychology depends ultimately on being able to fit various instances of our behavior together in such a way that others have confidence in the relative stability of our reactions in new situations involving both themselves and the shared environment. Our psychological self-ascriptions play an integral role in promoting and maintaining this confidence in our stability. But this is not just because they are verbal bits of behavior that fit, when all goes well, into the overall picture others form of us; it is, more importantly, because a facility with psychological self-ascription demonstrates to others that we understand just how much the stability and coherence in our behavior depends on how we govern ourselves. Hence, a general failure in our first-person authority would signal to others that we do not understand what it is to be a responsible agent, a lack of comprehension that invites doubts about our sanity or (minimal) cognitive competence—and probably rightly so.

This emphasis on responsible agency, on the ability to engage the expectations of others by making corresponding commitments through reliable first-person utterances, has led Bilgrami to take a

[29] *Brainstorms: Philosophical Essays on Mind and Psychology* (Cambridge: MIT, 1978); *The Intentional Stance* (Cambridge: MIT, 1987). Of course, there are ways and ways of "discounting" first-person claims. Others may have reason to suspect we are not competent in the language and, so, may give us credit for knowing our own intentional states despite the fact that we speak in idiosyncratic ways. They may consequently pursue a strategy of reinterpretation in order to preserve our first-person authority. I do not mean to rule this possibility out altogether; however, neither can the strategy of reinterpretation be elevated to the methodological pre-eminence suggested, for example, by a liberal interpretation of Davidson's "principle of charity." Preserving the other's rationality must be a holistic affair; linguistic competence is one of a number of interrelated competencies which, arguably, stand or stagger together. I elaborate on this point more fully in *The Meaning of Living Languages,* chapter IV.

more radical line on justifying the presumption of authoritative self-knowledge.[30] His argument, which I sketched in my introduction, is complex and difficult to recapitulate; but I take one of his grounding ideas to be as follows. If we value and so wish to preserve the structure of our various interactions—linguistic, social, political, moral—we cannot but engage with one another as responsible agents, agents that merit (for instance) praise and blame. But now we must be clear about what that decision (as it were) commits us to. It is an inextinguishable condition of being responsible that we have authoritative knowledge of our own mental states; for if we did not have such knowledge (generally speaking), we could not be held accountable for our acts since (in general) we could not be thought to know what we are doing. Knowing what we are doing involves knowing the intentional states that cause and give reason for our acts, but in the special way that implies *initiation* and, hence, *control.*

This is the crucial feature of Bilgrami's account. Knowing what we are doing in the first-person context—the context of *directive* agency—implies more than just knowing what first-order states are putatively causing our behavior. This is the limitation inherent in Wright's analysis: in presenting us as creatures simply assailed by a conscious awareness of our first-order states, we are unwittingly presented as utterly passive, not in control of our various thoughts and action, and so not able to take responsibility for them. To be viewed properly as agents, we must be construed instead as actively involved in forming, reviewing, revising, suppressing, and selectively acting on the first-order states we "know" about because we are the ones generating those very cognitive processes. The privilege of first-person knowledge is thus really more like the knowledge of a person driving a car as opposed to that of her passenger. The passenger may very well see where the driver is going, but still does not know in the immediate *executive* sense of the driver herself.[31] This analogy suggests a different, more profound kind of asymmetry than is captured by thinking about knowledge of mental states (literally or metaphorically) in terms of first- and third-person "points of view." For in making claims about one's own cognitive and emotional situation, one is making claims about a situation, both internal and external, that one has played (and continues to play) an active role in creating and maintaining. Hence, this is the kind of asymmetry which would make sense of the doctrine of first-person authority, even in the face of oc-

[30] See Bilgrami, "Self-knowledge and Resentment," especially sections IV and V.

[31] I shall pursue this point in much greater detail below. Heil has made a similar point in *op. cit.*, p. 248.

casional (perhaps even, systematic) error in particular first-person utterances.

The question, then, is how best to elaborate this line of thinking. The first step is to realize that we must give up a fiction to which many philosophers and psychologists have clung, even while rejecting cruder versions of the traditional picture of mind. It is that self-knowledge consists in coming to know (perhaps via theoretical mediation) a collection of facts: facts about one's own thoughts, feelings, intentions, and so on. Such accounts, whether externalist or internalist, naturally incorporate the structural hierarchy of intentional states I questioned above. For the underlying idea has been that one is in (or has) certain first-order states; one then forms second-order beliefs about these states; and finally, because of these second-order beliefs, one has the option to report one's cognitive situation to others. Of course, since these first-order states are the very states that cause one's behavior, one can be a pretty good predictor of one's future behavior to the extent that the second-order beliefs one expresses are relatively accurate. I shall call this the *reporter-predictor model* of authoritative self-knowledge.

In place of this, I propose we move toward an *agency model* of authoritative self-knowledge, which I shall characterize according to three distinct, but interrelated theses: (1) intentional states are dispositional; (2) self-ascriptions are, in many cases, commissive; and (3) first-person authority is an acquired capacity instilled in us, and preserved by us, only to the degree that we act, and expect others to act, as responsible self-directed agents. Since this is a presupposition of folk psychology, it is precisely in learning to participate in this practice that we learn how to think and act like authoritative agents.

*(1) Intentional states are dispositional.* Wright remarks that "one of the most basic of philosophical puzzles about intentional states is that they seem to straddle two paradigms: the paradigm of sensation and other avowable phenomena of consciousness, on the one hand, and, on the other, the paradigm of psychological characteristics that, like irritability or modesty, are properly conceived as dispositional and give rise to no phenomenon of avowal" (*op. cit.*, p. 631). This, of course, is what leads to the troublesome epistemology of intentional states. On the one hand, we do seem often to avow our intentional states fairly unreflectively, almost as if there were a way it characteristically feels to believe or desire that *P*; on the other, there are good reasons to suspect, with Ryle, that such avowals do not emanate from some determinate underlying state that both feels a certain way to us and causes in addition the various thoughts and actions characteris-

tic of believing that *P*. For Ryle, the feelings, thoughts, and actions themselves constitute a pattern of response we call "believing that *P*." So, if we avow the belief that *P*, it is because we recognize ourselves, just as others might recognize us, to act, think, and feel in ways that instantiate this pattern. We are not reporting, so not with special authority, on any underlying state at all.[32]

For the dispositionalist, this casts the problem of first-person authority into high relief. If we agree with Wright that our intentional practices persist only because there is some real authority attaching to our first-person claims as a result of others' being able to rely on them to predict our actions, how do we explain our prescience? That is, how do we explain our ability to ascribe dispositional states to ourselves prior to our manifesting the very patterns of behavior that underwrite such ascriptions? Are we not driven back to an "underlying state" view of dispositional claims: they are true just in case there is some determinate state persisting in mind, somehow detectable by us, and finally responsible for the characteristic "outward manifestations" of that state?

There is an alternative, but it requires adopting a more radical view of intentional states than even Ryle proposed (and concomitantly on the conditions of their avowal). The view I propose involves putting special emphasis on our own *agency* by recognizing that we are *actors* as well as observers and so can be good, even excellent, "predictors" of our future behavior because *we* have the power to make these "predictions" come true. Put simply, we are able to *ensure* a fit between the psychological profile we create of ourselves in first-person utterances and the acts our self-attributed intentional states are meant to predict and explain simply by adjusting our actions in appropriate ways. Thus, because we do not just wait to see if our actions make sense in light of intentional self-attributions, but rather *make* them make sense, the tale we tell of ourselves from the intentional stance is importantly unlike the tale we tell of other people (or even of other things). I cannot make it the case that you behave in ways coherent with what I say you hope, desire, or fear any more than I can make it the case that the world is a certain way by announcing how (I think) it is; but I can and do govern my own actions in ways that fit with the claims I make about myself. If so-called "knowledge" of our own minds thus consists largely of claims we have both made and acted in light of, it is no surprise that such "knowledge" is peculiarly authoritative.

[32] Ryle, chapter v: "Dispositions and Occurrences," especially section (2), "The Logic of Dispositional Statements."

We can see how this takes Ryle's analysis of psychological dispositions a step further. He claimed that we should not view the various feelings, thoughts, and actions that characterize a dispositional state as themselves emanating from some prior underlying state to which the agent herself has privileged knowledge-bestowing access. I add to this that the relevant bundle of feelings, thoughts, and actions come together in part because of the agent's own interests, reflections, and efforts in producing that bundle. That is, individuals *work* to make intentional characterizations true of themselves, especially when these are characterizations that they themselves have given.[33] Consequently, our "current-state" intentional self-characterizations have forward-looking truth conditions, or "satisfaction" conditions: they become increasing appropriate characterizations as these conditions are increasingly met. Agents have an obvious and particular ability to make such characterizations true of themselves. Thus, they have authoritative "self-knowledge."

*(2) Self-ascriptions are, in cases of "current state" ordinary avowals, commissive.* This feature of the agency model of self-knowledge has been implied by the foregoing: if so-called authoritative "knowledge" of our own minds consists significantly of claims we make about ourselves and then act in light of, such claims are analogous to promises in as much as they are *commissive,* not descriptive. The immediacy of their avowal is thus like the immediacy of making promises. Pace Wright, there is no internal state we are immediately detecting, even though we may be moved "subcognitively" to make such commitments—moved, in some sense, because of how we feel inclined to think and act. Nevertheless, we use our intentional self-ascriptions to do more than indicate how we feel inclined to think and act; for in using them, we become (even more) inclined to think and act in keeping with them simply *because* we have made such claims about ourselves. This is why our first-person ascriptions are, like promises, genuinely commissive. All the same, they must not be completely assimilated to promises, for they lack the specificity of promises as well as their moral dimension. I shall say more about these differences to bring out the notion of commitment I have in mind.

In the first place, our first-person psychological claims do not commit us to act in any specific way as promises do, but only to act in some way among others that makes sense in light of all that we have said about ourselves. Such ways will be many and varied, the more so

[33] I have explored these ideas more fully in "Psychological Dispositions: Revising the Philosophical Stereotype," MS (1995), written in collaboration with Eric Schwitzgebel at the University of California/Berkeley.

because they include the possibility of adjusting the psychological tale we tell of ourselves. So, for example, we can simply announce that we have changed our minds if we fail, or intend to fail, to act as people have been led to expect by our previous first-person utterances. Still, knowing when to make such announcements is part and parcel of what it takes to know ourselves. I stress this difference between nonspecific, yet genuine commitments and the more specific commitments of promises because of the following objection: it may seem that in proposing that many of our first-person psychological claims have a commissive element I have turned such claims into *motives* for behavior. But surely it is just crazy to suppose I always or, even, generally act in the way I do because I have made certain *claims* about my beliefs and desires. My beliefs and desires explain my actions, but my *saying* what I believe and desire is not my reason for acting as I do. My answer is that this is to adopt too narrow a vision of what we are committed to in saying what we believe, desire, hope, and so on. While it is true that I have no *added* reason to act as I do if what I do fits with the claims I have made about myself, I do have reason to change, modify, or add to the claims I make about myself when I recognize that what I do, or am about to do, does not fit very well with claims I have previously made. Since these verbal elaborations are actions, too, my (previous) words do give me reasons to act as I do.

In the second place, while I maintain that our first-person psychological claims have a normative quality, I do not claim that they carry any kind of moral imperative in the way that promises do. If we do not act in ways that make sense given what we say, even to the extent of modifying or superseding earlier claims at appropriate moments, we are not morally sanctionable; but we shall certainly be judged to be wrong about ourselves. This is not very significant in itself, though we often balk at having our authority questioned and tend to argue the point. Earlier I hinted at why this might be so. Being wrong about ourselves too much of the time affects our status as authoritative agents. At the extreme, the consequences of a general loss of authority, whether for good or bad reasons, are dire indeed. They involve various sorts of disenfranchisement—social, political, economic, legal, moral, and so on. An individual so treated becomes a patient rather than an agent, one whose behavior is first figuratively and then literally taken out of her control.

A further objection may be made to this commissive account of first-person avowals. Someone might argue that whatever plausibility it has derives from the fact that we have generalized from the wrong

sorts of cases. For we have focused on claims that an agent makes about herself *to others*. What about judgments that are made by an agent but unvoiced? Are we really to think that in these cases an agent is making commitments to herself to act, speak, and operate in ways that accord with such claims? What would be the point of such commitments? How could they be binding in any way?

Here, I think, we must simply bite the bullet, as we do in the limiting case of supposing someone can make a promise to herself. The phenomena of self-reflection and self-report are in no way special for being unvoiced. Whether our judgments about our own intentional states are expressed publicly or not, they inherit their commissive quality from the training we have received in the public language of folk psychology. For in learning to use this language to apply to ourselves, we have also been taught to act and think in ways that demonstrate responsible (self-knowing) agency. And this depends, in part, on developing self-regulating habits of mind that use our moments of self-judgment to keep our various mental and physical activities in line with what we are taught makes sense. In the course of development, we thus become to a very great extent self-programmers: we learn to use our intentional self-ascriptions to instill or reinforce tendencies and inclinations that fit with these ascriptions, even though such tendencies and inclinations may at best have been only nascent at the time we first made the judgments.

Dennett, I think, is among those who would be sympathetic to this idea of how we regulate ourselves. In describing the psychological consequences of our being linguistic creatures, he writes:

> Language *enables* us to formulate highly specific desires, but it also *forces* us on occasion to commit ourselves to desires altogether more stringent in their conditions of satisfaction than anything we would otherwise have any reason to endeavor to satisfy. Since in order to get what you want you often have to say what you want, and since you cannot say what you want without saying something more specific than you antecedently mean, you often end up giving others evidence—the very best of evidence, your unextorted word—that you desire things or states of affairs far more particular than would satisfy you—or better, than what would have satisfied you; for once you have declared, being a man of your word, you *acquire* an interest in satisfying exactly the desire you declared and no other.[34]

*(3) First-person authority is an acquired capacity.* For me, the interesting questions begin here: How does our training in language (particularly in the language of folk psychology) along with our training in

[34] *The Intentional Stance*, p. 20.

acceptable forms of behavior turn us into "men of our word"—that is, turn us into agents who (authoritatively) know our own minds because we take the responsibility for following through on claims we make about ourselves? Why is it necessary for us to develop the interrelated linguistic and nonlinguistic competencies involved in treating our first-person psychological claims as claims that commit us to act in various nonspecific but sense-making ways?

I suggest a short answer to the latter question, saving its elaboration for another occasion: it is necessary for ensuring that we shall act in ways that create and maintain a stable social environment—that is, one in which each of us can, from the intentional stance, predict and explain what others will do with reasonable success.

As to the former question, I suggest we need to look at the practice of folk psychology in a rather different way than is currently accepted in the philosophy of mind and (derivatively) in certain burgeoning areas of cognitive and developmental psychology. Within these contexts, the common assumption is that folk psychology is a fairly systematic, but theoretically unsophisticated practice developed over generations to explain and predict people's behavior. Many claim that, like folk physics, it is not immune to advancing scientific knowledge and will no doubt be either significantly altered or replaced altogether by a respectable scientific picture of the mind/brain. In philosophy, there is fairly active debate about just how much of folk psychology will remain come the new dawn: Will we even need the theoretical concepts of belief and desire, let alone the fuzzier notions of traits like honesty, steadfastness, cowardice, and unreliability? There is also active (and related) debate about why folk psychology (unlike folk physics) has persisted so long in its relatively naive and haphazard form. Some say that, unlike folk physics, folk psychology must do a pretty good job of limning the real causal structure of the mind; others say, understanding mental causation is just a very hard problem, so developing the new science has taken (and will take) us a long time (though "neurophilosophers" do think we are getting much closer indeed).

My own view is that these debates miss something profoundly important about the practice of folk psychology. We can agree that folk psychology, like folk physics, is in the business of articulating generalizations about how the objects in its domain (people) behave under various conditions. But folk-psychological generalizations, unlike folk-physical generalizations, are first and foremost *normative.* They are also descriptive, but unlike folk-physical generalizations, they are descriptive *because* they are normative. That is, folk psychology is suc-

cessful primarily because we become the kind of agent it successfully predicts and explains by molding ourselves and others to suit its generalizations. Thus, a primary use of folk psychology is normative—to teach and also to remind each other to be what we think of as competent intentional agents. This means its generalizations could not be falsifiable in anything like the way the generalizations of nonintentional scientific (or even protoscientific) theories are. And, indeed, our folk-psychological generalizations are often honored as much in the breech as they are in the observance—primarily through practices of correction and critical reflection. This explains why folk psychology has not withered away in the face of what might seem a constant supply of "recalcitrant evidence."

Of course, if I am right to claim that our common-sense psychology is a normative practice we learn to engage in for the purpose of becoming responsible, understandable agents, the norms we use to circumscribe our behavior need not be immutable. Indeed, they are likely to be susceptible to the very kinds of pressures that cause norms in general to evolve.[35] While these kinds of pressures are various, some at least come from currently dominant scientific theories and practices. This leads to a different way of thinking through the discomfort many have felt with the claim made by some members of the philosophical avant-garde that common-sense psychology is a moribund, quasiscientific practice which ought to be replaced by a respectable scientific picture of mental causation. For though its detractors would like to argue the point, the project of replacing folk psychology with a more "scientific" way of understanding ourselves is not in principle doomed to failure. But, pace the advocates of such reform, I do not think such changes would be demanded by so-called "facts" about how our minds work. Minds are as much made as discovered.

This is a substantial thesis, and it needs defense. In order to provide such defense, we must not ignore the developmental evidence; far from it. But what evidence there is could be re-examined to try and discern the interactive processes whereby children come to fashion themselves in the folk-psychological image of the intentional agent. Then, perhaps, we can use this knowledge to explore the developmental and everyday implications of adopting a conceptual framework in discussions of mind which downplays first-person authority. Does this mean I am some kind of antirealist about the mind?

[35] I have already dealt with this phenomenon at some length in relation to linguistic norms in my doctoral dissertation, *The Meaning of Living Languages*, MS, University of Toronto (1990). The model I develop there can be easily extended.

Only if a commitment to realism requires embracing a certain scientific naiveté. If it is true that we mould ourselves to fit the explanatory and predictive framework of the day, we ought to be concerned (as essentialists refuse to be) with the real possibility that first-person authority is precisely the kind of human trait that could be diminished or lost in our scientific revisions.[36] I mean this not as a slight to science, but as a reminder that humanism has its own enjoinders, enjoinders that can, in fact, promote good and reflective science. In this case, we are enjoined to consider how our commitment to first-person authority ramifies, as Bilgrami argues, through our various spheres of interpersonal contact—private, medical, legal, social, political, and so on. Moreover, it is in reflecting on these things that we remind ourselves of why such authority is worth preserving and why we should encourage practices (whatever they may be) that support it.

## III. CONCLUSION

In the end, the account of self-knowledge that I propose has certain important affinities with Ryle's, despite his behaviorist leanings. For instance, we agree in de-emphasizing the self-attributive use of psychological terms to report on one's mental states rather than to express, in a shared language, one's propensities and inclinations to act in further verbal and nonverbal ways. What this does not preclude, of course, is the activity of using psychological terms self-reflectively—that is, of joining with others in critically reflecting on one's apparent propensities and inclinations to act in various ways in

[36] As Bilgrami warns, there has always been in some circles, and is perhaps now on the increase, a "psychiatric model" of human behavior that replaces the structuring ideal of the responsible agent with the notion of a treatable patient—one whose affective responses are debilitating and best controlled by therapy or (increasingly) medication. Of course, in "fixing" the patient, little heed may be paid to the coherence of her responses to environmental conditions. If the person is indeed responding in a coherent way, it may be wondered how adjusting such responses affects her long-term ability to understand her own experiences as manifestations of a stable and coherent persona. That is, if the person is increasingly directed to attend to her current feelings with an eye to alleviating them, how is she to use her own experiences to build and modify her understanding of the nature (and rationality) of human response to a complex world? Relieved of the need to understand whether her responses are generalizable because they make sense under particular circumstances, she is relieved also of the motivation for challenging those circumstances which give rise to her current experiences. These worries are highly speculative, to be sure, and it may sound as though they stem from incipient science anxiety; but that, at least, would be wrong. Neuroscience has much to tell us about how the mind/brain functions—what neurotransmitters control affective and motor responses, what areas of the brain underwrite what capacities, and so forth. Cognitive and social psychology have equally much to tell us about our higher level capacities for organization of information and response to the world. My only concern is that, amidst all of these exciting scientific discoveries, the philosophical dimension of questions raised and considered in these contexts not be forgotten or ignored.

light of what one has led people to expect, in large part, by one's own prior verbalizations. Knowledge in this sense is backward looking: it depends on correctly or sensibly characterizing the *fit* between an individual's previous utterances and her nonlinguistic acts, whether or not that individual is oneself. I agree with Ryle that, in this particular self-critical activity, one has no special authority about the veracity of one's claims. At best, one has a contingent, residual authority derived from being around oneself all the time and so possibly, in a given situation, knowing more relevant things about what one has said and how one has acted than others do.

Nevertheless, while Ryle was right to disparage the idea of what he called a "metaphysical iron curtain" separating knowledge of our own thoughts from knowledge of the thoughts of others, he was wrong to suppose that *privileged access* could be the only grounds for supposing there is in general a *principled* difference in the authority attaching to first-person psychological claims and the authority attaching to claims made from any other perspective. For as Ryle rightly noted, many (probably most of) such first-person claims are not reflective claims at all; nor are they merely "effects of the frames of mind in which they are used" (*op. cit.*, p. 175).[37] But whereas Ryle was content to treat "unstudied utterances" as expressions of the states of mind they consequently "disclose," my own view is that we use such utterances more actively to tell a constantly updated story about ourselves that we also act upon to make true. Our stories are told in an intentional idiom we use and have been taught to use for the purpose of making our behavior sensible, predictable, and understandable to others, as well as to ourselves. For in learning how to use this idiom, we also learn what it is to "have" beliefs, desires, and other intentional states: they are states that allow others to explain and predict our behavior; hence, they are states that, claimed by us, commit us to act in various ways, even if those ways include explaining why we have not acted, or shall not act, in the ways we have led others to expect.

[37] It should be noted that, in characterizing his "avowal theory" of unstudied talk, Ryle clearly pre-empts the objection that first-person claims about beliefs, desires, thoughts, and the rest are merely articulated grunts on par with other behavioral manifestations of particular frames of mind. As he says: "While careful observation of the subject's other behavior, such as his overt action, his hesitations, and his tears and laughter, may tell him much, this behavior is not *ex officio* made easy to witness, or easy to interpret. But speech is *ex officio* made to be heard and made to be construed. Learning to talk is learning to make oneself understood" (*op. cit.*, p. 176). Still, Ryle was prevented, in my view, from making anything further of this crucial distinction between verbal and nonverbal behavior because of his reductionist views about the nature of psychological states.

This actively constructive ingredient, missing from Ryle's analysis, makes clear why our first-person psychological claims underwrite judgments that we know our own minds, even though our first-person claims are not themselves expressions of true, justified beliefs about our own first-order intentional states. We know our own minds because we have been trained to take on the responsibility, as only cognitively and linguistically sophisticated agents can, for suiting our words to our deeds and our deeds to our words. Once we admit this—that "knowledge" of our own minds is peculiarly dependent on our role as agents—we can readily explain, as Ryle could not, the *principled* difference in authority attaching to first- and third-person psychological claims.

VICTORIA MCGEER

Vanderbilt University