CHAPTER ONE

The Image of Self-Knowledge

The question of the nature of first-person relations has not suffered from philosophical neglect in recent years. Perhaps unsurprisingly, attention has tended to concentrate on the particular relation of *knowledge*; and even more particularly, on the specifically first-person awareness we normally take ourselves to have of our own mental life. This chapter attempts to reorient some of our thinking about self-knowledge and place the more familiar epistemological questions in the context of wider self-other asymmetries which, when they receive attention at all, are normally discussed outside the context of the issues concerning self-knowledge. This task is really the concern of the book as a whole, but this first chapter seeks to establish three related points.

The first is simply the proper characterization of the basic difference between how a person may know his own mind and how he may

know the mind of another. In one guise or another, this is a familiar idea, and not only in philosophical discussion. However, the various services it has been pressed into, especially in the history of epistemology, have obscured the basic asymmetry and its rationale and freighted the idea of self-knowledge with a host of extraneous philosophical assumptions. For a long time, the problem of distinctively first-person awareness has led a kind of stepchild existence in philosophy, much less often investigated for its own sake than in the context of other problems, either concerning epistemological foundationalism and materialism, or, more recently, externalism about mental content and skepticism about meaning and its determinacy. This has contributed not only to a narrow view of the range and variety of first-person knowledge, but also to a distorting emphasis on various extreme and contentious claims about its nature and extent, which has deflected attention away from the basic differences that remain between knowledge of oneself and knowledge of others, even after the abandonment of anything resembling "introspective infallibility." The wider view of self-other asymmetries, however, within which any such specific claims of first-person authority must take their bearings, obliges us to ground the discussion as much in moral psychology as in epistemology.

A second concern will be to inquire how it is that philosophical accounts of self-knowledge often fail to account for (or sometimes even to describe) a specifically first-person phenomenon. Put somewhat less paradoxically, prominent accounts of self-knowledge often end up either describing something that could just as well be a third-person phenomenon, or transposing an essentially third-person situation to some kind of mental interior. The "internal theater" of Descartes (and Locke and Hume) and the long legacy of treating self-consciousness as a kind of inner perception is probably the most graphic expression of this approach, but the general tendency is broader than this.[1]

[1] The most continuous and consistent case against the Perceptual Model generally has been made by Sydney Shoemaker, beginning with *Self-Knowledge and Self-Identity* in 1963, and in a series of papers in the decades since then. See especially (1986),

Nothing especially first-personal is captured by transferring the situation of a spectator from the outside to the inside, nor by construing the person as having any kind of especially good theoretical access to his own mind. A theme throughout the book will be that the difficulties in properly characterizing the first-person position are not merely epistemological ones, and later chapters will take up the theme of characterizing first-person relations that are not based on third-person models and do not involve essentially alienated relations to the self.

A final purpose of this chapter is to show how the lingering influence of a Cartesian picture of introspection creates unwarranted skepticism about the very possibility of self-knowledge. For current purposes, we can see the "Cartesian picture" as combining a radical epistemological claim of infallibility with a characterization of this particular *mode* of awareness as a kind of internal perception. Hence the picture of the "inner eye," incapable of error. Both aspects of this picture have been subject to a good deal of criticism in the past century, but often against the background assumption that introspective access must be something that conforms to this picture if it is to be anything distinctive at all. Hence the fate of Cartesianism has been taken to be decisive for the very notion of first-person awareness, and recent philosophical work has been very creative in developing ways of describing the surface phenomena of first-person discourse that are deliberately deflationary of the claims of that discourse to be reporting any kind of genuine awareness. By contrast, I wish to defend a view of first-person awareness that sees it as both substantial, representing a genuine cognitive achievement, but which nonetheless breaks decisively with the Cartesian and empiricist legacy. As subsequent chapters will show, this entails not only rejection of the "inner eye" as applied to the mechanism of introspection, but an account of the general distortions of the purely theoretical or spectator's stance toward the self (both as expressed in philosophical accounts of introspection, and in the life of

(1988), (1990), (1991), and (1994). I have not tried to relate in detail the story I tell here to this body of work, but anyone familiar with it will know what I owe to it.

the self). Being the person whose mental life is brought to self-consciousness involves a stance of agency beyond that of being a kind of expert witness. Thus the discussion taken up here moves from the epistemology of introspection to a set of issues in the moral psychology of the first-person.

1.1 THE FORTUNES OF SELF-CONSCIOUSNESS: DESCARTES, FREUD, AND COGNITIVE SCIENCE

The legacy of Cartesianism has been decisive in the philosophy of mind not only in the positive influence it exerted in the centuries immediately following Descartes, but just as much in the force of its repudiation in the twentieth century. Nowhere is this clearer than in the question of the mind's access to itself and its operations. Recent philosophy typically rejects the picture of the mind as immediately transparent to itself, and then tacitly takes this rejection to be equivalent to rejecting the very idea of introspective access, thereby ceding the very concept of first-person awareness to its Cartesian interpretation. In this first section, I mean to trace some of the main outlines of this story, with the aim of disentangling the basic idea of first-person access from the Cartesian picture. One striking fact about this story is that although largely Cartesian assumptions about the mind's access to itself dominated both epistemology and philosophy of mind since the seventeenth century, it wasn't until the twentieth century that the problem of the person's "privileged access" to his own mental life was treated as a philosophical issue in its own right. Earlier, major figures within both empiricist and rationalist traditions could take for granted that there is nothing in the mind of which the person is not conscious, and that a person's knowledge of his own current mental states is both certain and infallible; in short, that the mind is "transparent" to itself. In the centuries since Descartes, the identification of the mental with consciousness was more often treated as a guiding assumption than as a positive thesis in philosophy, something only occasionally given

explicit formulation and defense.[2] And even when, in the twentieth century, the nature of "first-person authority" was identified as a philosophical issue of its own, the primary interest was not so much to investigate or defend the assumption itself as, rather, to give an account of *how* such privileged access was possible, and to find ways to accommodate certain assumptions of infallibility within the logic of first-person discourse.

As a result, it is often assumed today that the idea of philosophically important differences between self-knowledge and knowledge of others depends on maintaining a thesis of introspective infallibility in some form or other. And indeed, the contemporary rejection of this thesis has led various philosophers to reject the idea that there is anything philosophically distinctive about self-knowledge as a type of awareness.[3] This attitude is often supported by appeal either to psychoanalytic theory or, more commonly these days, to various results of experimental psychology and contemporary cognitive science, which seem to show that people's reports about their own reasoning processes are often unreliable and that various aspects of information processing are in principle inaccessible to consciousness.[4]

This development is in turn part of a curious and radical swing exhibited by recent philosophical thought about the mind. In this past century, philosophers have gone from thinking of the mind as totally open to introspection to doubting not just the reliability but the very reality of introspection. Early in the century, for instance, Freud often complained of the opposition he encountered to the very idea of unconscious mental processes from philosophers who simply *identified*

[2] This is not to suggest that the idea of the mind's perfect access to itself was entirely without its dissenters throughout this period. For the explicit denial of this claim, one need go no further than Arnauld's objections to the Meditations (Objections IV, Haldane and Ross, p. 92). See also Objections VI, p. 235.

[3] Chapter 4 of Churchland (1984) is guided by this assumption. See also Rorty (1982) for the claim that "our knowledge of what we are like on the inside is no more 'direct' or 'intuitive' than our knowledge of what things are like in the 'external world'" (pp. 330–31).

[4] See, for example, Nisbett and Wilson (1977), and Nisbett and Ross (1980).

"mental event" with "that which is immediately present to consciousness" (Freud 1915). Such an identification was part of philosophical common sense on both sides of the Atlantic for a long time. Nowadays, however, philosophers are apt to express doubt whether *anything* of psychological significance is an object of introspective awareness. Daniel Dennett once expressed this thought in the following terms: "The control of reflexes in man is subconscious, as are the stages of perceptual analysis, and in fact all information *processing*. We are not aware of the processes at all (as one might, with suitable incisions and mirrors, be aware of one's digestive processes). . . . As Lashley says, 'No activity of mind is ever conscious." (1969, p. 128)

Clearly, *some* kind of reconception of the mind and its access to itself is represented in this development, but to take the apparent scientific opposition at face value would be to miss what is distinctive about first-person access. That is, it would be a mistake to see this theoretical development in terms of a single, stable conception of the mind, with respect to which philosophers and psychologists have somehow gone from seeing all its activities as transparent to itself to seeing virtually none of its activities as belonging to consciousness at all. It is not the same sense of 'mental event' that was once thought to be intrinsically conscious and is now seen to be no more conscious than the breakdown of sugar in the body. In particular, it matters that the unconscious thoughts Freud postulated were understood by him to be the very sort of thing that could, in principle, be brought to consciousness: the familiar states of mind of belief, emotion, and desire. By contrast, with respect to psychological processes identified at the "subpersonal" or "computational" level of description, it is doubtful that we have any sense of what could even *count* as introspective awareness. For this reason, Dennett's comparison with observing the processes of digestion actually *under*states how "introspectively unavailable" the "activities of the mind" would have to be on the computational model assumed in this passage. For with respect to digestion we would at least have some *idea* of what the scene might look like "on the inside," whereas activities subpersonally described provide no such idea, and

hence no idea of a *failure* of such access, either.[5] Whatever introspective access is, our understanding of it will have to make sense of its conceptual dependence on the level of commonsense psychological description. For the object of first-person awareness (on *any* account of it) is not all of psychological life, but primarily the states of mind identified under the categories of what is sometimes called "folk psychology": the hopes and fears, pains and experiences we relate to each other in daily life, and not states or processes defined either neurologically or computationally.[6]

This conceptual difference in the *kinds* of states of mind in question (and the level of description appropriate to them) is closely related to a difference in the ordinary *importance* of the availability to consciousness of various psychological phenomena. Here the distance between Freud and Lashley is as great as that between Freud and his earlier philosophical opponents. For in his practice, Freud was concerned to restore something to consciousness, which was an ordinary, if incomplete and insecure, possession of people, and one that was understood to be crucial to the conduct of life. Blindness or ignorance *here*, unlike, say, with regard to the facts about the internal processing of visual stimuli, was understood to be disturbing and debilitating. By contrast, ignorance of the psychological facts of the sort alluded to by Dennett is the normal case for all of us, and those who do know something here do not know it in anything like the way ordinary people take

[5] Recently, John Searle (1992) has taken the inaccessibility to consciousness to be a reason to deny that such subpersonal states and processes posited by cognitive science have any right to be considered *psychological* in any serious sense. While I agree that the distinction of levels of description is extremely important, I don't think our concept of the mental is as rigidly defined as Searle's denial would seem to require.

[6] Christopher Peacocke has emphasized the importance of this distinction for any account of self-knowledge, and in "Conscious Attitudes and Self-Knowledge," he means to reject "the position of someone who says that there is never a personal-level, causal, reason-giving explanation of why a thinker has the belief that he has a certain belief, in normal cases" (1998, p. 77). In my own account, the relation between self-knowledge, the personal-level, and reason-giving will be elaborated in later chapters.

themselves to know about their own thoughts and feelings. Thus, another task for understanding introspective availability is to understand why it should have any importance to ordinary rationality and personhood. It's natural to take the normal importance of self-consciousness for granted and to assume we understand *what* its importance is, but it should not be so obvious once we reflect that mental phenomena may be identified in many different ways (neurologically, computationally, in everyday terms), and that many perfectly rational and adaptive processes neither require nor tolerate self-conscious monitoring for their proper functioning. The normal importance of self-knowledge in a person's life will have to be understood as dependent on the level of description provided by the concepts of ordinary "folkish" psychological discourse.

The problematics of self-knowledge for both commonsense and Freudian thought employ a conception of the mental that is distinct from both the Cartesian picture and that of contemporary cognitive science, as different as these models are from each other. Freud argued against the philosophical identification of the mental with the immediate presentations of consciousness, but he did not deny that there *is* such a thing as ordinary introspective awareness. And he took it to be important to mental health that a person's beliefs and so on should normally be available to him in this way. His claims about the incompleteness and fallibility of this mode of awareness are thus in sharper opposition to the radical Cartesian claims of the mind's transparency to itself than they are to much of the commonsense understanding of self-knowledge (a discourse that allows for the possibility of *difficulty* and failure here, and doesn't contain terms like 'introspective infallibility'). This is not to suggest that there is *no* conflict between Freud and commonsense views of the mental, but it is important to see that the Cartesian picture is something theoretically distinct from either of them. For Descartes' picture assimilates a great range of psychological phenomena to something like the status of episodes of consciousness [*cogitatio*], thus classifying even the operations of will and judgment

together with sensations, mental images, and passing thoughts.[7] In this way, grounding the general category of the mental in the paradigm of the experiential and the episodic lends a misleading plausibility to the characteristic Cartesian claims of introspective infallibility and self-intimation. For while it is indeed difficult to conceive of the possibility of, for example, intense pain of which one is utterly unaware (a possibility which even Freud himself, for instance, never countenanced), this is so for reasons quite specific to the special case of pain and does not carry over to motives, moods, beliefs, and the rest of what we commonly think of as belonging to the psychological. The general Cartesian category of the inner is something with a particular philosophical motivation, and indeed a good part of this motivation lies in the effort to identify an epistemological foundation that is precisely *not* prone to the sorts of gaps and errors that belong to our judgments about the external world.

The dependence of the Cartesian picture on such specific epistemological motivation provides all the more reason not to identify it with the general problematics of self-knowledge. For independent of this picture there remains a set of basic asymmetries between self-knowledge and the knowledge of others, which point to a different set of philosophical questions concerning how self-relations necessarily differ from relations with others. There are two basic categories of psychological state to which the ordinary assumption of "privileged access" is meant to apply: occurrent states such as sensations and passing thoughts, and various standing attitudes of the person, such as beliefs, emotional attitudes, and intentions. (I will have comparatively little to

[7] This claim of the revisionary distinctiveness of the Cartesian category of the mental and its distance from "common sense," is not itself free from controversy; but in support of such a view, see, for example, Kenny (1968). Such an assimilation of diverse psychological kinds is not unique to Descartes, of course. The doctrine of "ideas" in Modern Philosophy generally did not make our contemporary philosophical distinction between, e.g., images and thoughts, and this made it easier to think of all mental life as immediately present to consciousness.

say here about the case of sensations, which I believe raises issues for self-knowledge quite different from the case of attitudes of various kinds.) The type of access we ordinarily take ourselves to have here is special in at least two basic ways. First, a person can know of his belief or feeling without observing his behavior, or indeed without appealing to evidence of any kind at all. And second, rather than this nonreliance on evidence casting doubt on the reliability of such reports, judgments made in this way seem to enjoy a particular epistemic privilege not accorded corresponding third-person judgments that *do* base themselves on evidence. For now, we need not concern ourselves with just how strong this epistemic privilege is supposed to be, for example, whether such judgments are "incorrigible" or not. Suffice it to say that they are taken to have a good prima facie claim to truth which may be overruled only in special cases. The important point is that these are taken to be genuine judgments, expressive of knowledge, which are made without reliance on "external" observation. This will need explaining, even if one is inclined to dismiss the larger claims of infallibility or incorrigibility.

The claim that introspective awareness is not inferred from observational evidence is what is usually intended by the claim that it is "immediate." As a claim about the *mode* of awareness, this just means that such judgments are not inferred from anything epistemically more basic. Beyond that, immediacy does not entail anything about the epistemic authority of the judgments. Judgments with this immediacy need not in general enjoy any special kind of certainty, as compared with some other judgments that may base themselves on observation and inference. In the case of knowledge of oneself, it is particularly clear that the judgments that may be immediate in this sense concern a subject matter (i.e., a certain person's mental life) about which judgments are made in other ways as well. A person may report on his own emotional state introspectively, but at the same time he recognizes that other people come to their own conclusions about this same state of his in quite different ways. And the person himself may on occasion employ such third-person evidence in learning about his true emo-

tional state. Thus there are both immediate judgments and evidence-mediated judgments he may make about the same question, and for all that has been said so far, it could well be that immediacy per se does not confer any greater reliability or freedom from error on a set of judgments. As with perceptual illusion, it could be that immediate awareness is available in an area where such judgments are nonetheless prone to characteristic errors for quite independent reasons. It could be that, for a range of psychological states, immediate introspective awareness is a less reliable guide, subject to characteristic errors of its own, than are the observation-based judgments of others. The claim of immediacy then, is a claim about the specific *manner* of first-person access and should therefore be kept separate from any epistemic claims of infallibility or introspective certainty.

In addition, immediacy is to be understood as a wholly negative claim about the mode of first-person access, that is, awareness that is not inferred from anything more basic. Much of our ordinary perceptual awareness is also taken to be immediate in this sense, and this particular model of immediacy, enshrined in the etymological connotations of the word 'introspection' itself, has irresistibly suggested to many philosophers that introspective awareness is immediate because it is itself a form of perception, a kind of "inward glance." However, it is important to note at the outset that identifying introspection with a kind of perception is a substantive philosophical *interpretation* of immediacy and is not simply equivalent to it. The basic concept of first-person awareness that we are trying to capture is that of awareness that is not based on evidence, behavioral or otherwise. This basic concept of immediacy is itself not wholly free of controversy, of course, but the perceptual model of "introspective" or "first-person" access is an additional substantive thesis, which, while not without its contemporary defenders, has been subject to sustained criticism in this past century.[8] What we are identifying as the full-blown Cartesian picture

[8] In addition to the papers by Shoemaker mentioned in note 1, arguments against construing introspection as a form of perception may be found in Wittgenstein (1956),

of introspection combines both the strong epistemic claims of infallibility and self-intimation, and the characteristic perceptual model of just what manner of awareness introspection is.[9] Independently of this model, we can characterize a set of basic asymmetries between knowledge of oneself and knowledge of others that survives philosophical attempts either to dismiss it or to explain it as a consequence of the merely spectatorial advantages of the first-person point of view.

1.2 THE POSSIBILITY OF SELF-KNOWLEDGE: INTROSPECTION, PERCEPTION, AND DEFLATION

What remains before us, then, is a basic asymmetry between first-person and third-person relations. A person can make reliable psychological ascriptions to himself immediately, without needing to observe what he says and does. And this capacity lies in the nature of the first-person position itself; it is not a kind of access he may have to the mind of another person. Compared with the traditional Cartesian doctrine of

---

Sellars (1962, p. 33, and 1956, p. 178), Davidson (1987), and Evans (1982). In recent years, the main defender of seeing first-person awareness as a form of internal perception has been David Armstrong (1968, 1984).

[9] Was Descartes himself a Cartesian in the sense just defined? This is a somewhat vexed question. There is certainly a preponderance of evidence that he was committed to some versions of the doctrines of both infallibility and self-intimation. See, for example, M. Wilson (1978, in particular p. 151). However, Descartes himself is not entirely consistent about this, and wavers about the claim of Self-Intimation in particular. In a letter to Gibieuf (January 19, 1642), he writes: "But I do not deny that there can be in the soul or the body many properties of which I have no ideas; I only deny that there are any which are inconsistent with the ideas that I do have" (Kenny 1970, p. 125). And in the *Discourse on the Method*, he even employs a version of the "distinct existences" argument against Self-Intimation, nowadays associated with David Armstrong: "Many are themselves ignorant of their beliefs. For since the act of thought by which we believe a thing is different from that by which we know that we believe it, the one often exists without the other" (Descartes, vol. 1, p. 95).

As for the other half of the Cartesian picture, the perceptual model of introspection, it is probably correct that he assumed some version of it, but its explicit formulation is much clearer in other Modern philosophers, for instance, Locke in Book II of the *Essay* (1690).

introspective infallibility, this is a relatively modest characterization of the "privileged access" a person has to his own mental life, but it is hardly either psychologically or philosophically innocuous. Various aspects of introspection's claim to either completeness or reliability surely are challenged by Freud's theory of the unconscious as well as by contemporary cognitive psychology. But perhaps more pressing than the question about the epistemic completeness or reliability of introspection are philosophical questions concerning how there could even *be* such a thing as this capacity, however imperfect its deliverances. How is it possible for there to be knowledge of some contingent matter of fact (e.g., the facts about what I believe or hope for) that is not based on observation of some kind? And in what sense is this knowledge supposed to be essentially or exclusively first-personal?

To many philosophers, these worries have suggested that so-called introspective judgment cannot be construed as the genuine "detection" of some independent psychological fact, and that the logic of "avowals" must be given an analysis that explains away their appearance as expressive of first-person *judgments*. There are a number of different forms such a "deflationary" account may take, but I want first to say something about what motivates the search for an account of this type.

As suggested, sometimes doubt is cast on the very possibility of introspective self-knowledge by the lingering assumption of some kind of perceptual model for it. One may think that if what we are doing in introspection really involves the detection of some independently obtaining state of affairs, then it could only be by means of some kind of perception. At this point, we encounter various difficulties in applying this picture, and instead of challenging the picture, philosophers may be more prepared to deny the substantiality of introspection itself. The first such difficulty is the original embarrassment of the "inner eye" and the concern that it cannot be cashed out as anything other than a misleading metaphor. There is no perceptual organ of introspection, in anything like the way there are identifiable organs of sight and hearing and the like. Further, something like a person's sensation of red is not to be analyzed into an independent object accompanied by an act of perceiving it. Aside from familiar ontological

problems with the reification of sense data, and regress problems with the idea of "the perception of an appearance," there simply doesn't even seem to *be* any "appearance" or perceptual presentation of one's belief or sensation that would be the experiential basis for the quasi-perceptual judgment, for example, that one has a headache, or believes that Wagner died happy. While "representationalism" is a controversial thesis about the ordinary perception of objects in the world, on nobody's view is the awareness of one's headache mediated by an appearance *of* the headache. And in the case of attitudes like belief, there is simply nothing quasi-experiential in the offing to begin with. There is nothing it is *like* to have the belief that Wagner died happy or to be introspectively aware that this is one's belief, and that difference does not sit well with the perceptual analogy, even if the problems it encounters with respect to sensory states could be solved.

Rather different problems with the analogy arise from recent considerations in the theory of mental content. Many philosophers identify what would be linguistically expressed as a simple single belief with a state of the person satisfying a complex functional role, involving a vast array of potential inference patterns, conceptual commitments, and dispositions to behave. On such a view, the simple belief that Wagner died happy is constituted by a host of inferential commitments concerning related matters (about Wagner, death, happiness, and much else) and the truth of various counterfactuals. How, then, one may ask, could all of *this* be presented to my immediate inner perception when I am aware of what I believe about Wagner? I don't even know what "all" of this *is*; in fact, I may be explicitly aware of hardly any of it, and yet the belief I am supposedly aware of is constituted by nothing less than all this (and not by any graspable mental image, for instance). If a person is indeed aware of his own belief, it is not by being somehow perceptually presented with anything of this complexity, for there is no such presentation.[10]

[10] The doctrine known as holism comes in many varieties, but it should be noted that the difficulty for the perceptual model given here depends only on a quite modest claim

A closely related problem with the model is posed by "externalism" about mental content, the claim that what a thought or belief is *about* may be determined by relations the person bears to various environmental factors of which he may have no knowledge at all. For both functionalism and externalism, then, the identity of a thought is constituted by various relational properties. Paul Boghossian (1989) takes this relational feature of such views to present a prima facie case for skepticism about the very possibility of introspective awareness of one's thoughts. On this view, if some form of externalism were true, then from within the first-person perspective one would be in a position analogous to that of someone inspecting the intrinsic features of a coin to determine its monetary value, but where the value of the coin is wholly determined by, say, where it was minted and is not indicated in any way on the face of the coin (p. 16). It is such relational facts that determine the value of the coin, and these are not part of its observable features. All that introspection can deliver is awareness of the intrinsic properties of a thought-token, and hence, on the assumption of externalism, such awareness is really no better than blindness. Boghossian himself is clear that he does not mean to endorse skepticism about self-knowledge, but he does nonetheless take the difficulties presented by the assumption of externalism to be quite real ones.

I have only presented a sketch of this skeptical argument here, and I will not be concerned with the various problems that may be found with this style of reasoning.[11] Here I only want to point out how in both cases (functionalism and externalism) the appearance of a skeptical threat depends on assuming the appropriateness of the perceptual

---

about the complex constitution of a given belief state. The truth of holism would, of course, only make things worse for the model.

[11] Boghossian's paper has contributed to an explosion of literature on this subject recently. A bit earlier, both Burge (1988) and Davidson (1987) developed accounts aimed at reconciling externalism and ordinary self-knowledge. Two recent collections of papers are Ludlow and Martin (1998), and Wright, Smith, and Macdonald (1998). In addition, see Falvey and Owens (1994), as well as recent work by Ebbs (1996 and 1997).

model of introspection. This is especially clear in Boghossian's original presentation, as can be seen from his comparison of the situation of introspection with the example of the coins. For the comparison to work in this argument, we would have to understand the ordinary case of awareness of our own belief (e.g., that Wagner died happy) as proceeding via some quasi-perceptual presentation, which we then need to interpret as having a certain representational content. But this is manifestly not a person's relation to his own thoughts and beliefs, however mental content is determined. That would be just as inapt as suggesting that the way that I know that I'm thinking of my mother, rather than my aunt her twin sister, is due to the fact that she wears her hair differently. The idea that intrinsic or phenomenal features of some mental experience do not tell the person what his experience is about has been a familiar idea at least since Wittgenstein and was part of his original case against the perceptual model. And when he asks, "How do I know that I am imagining King's College on fire, and not another one just like it?," his point is not to suggest any doubt about what he is imagining, but to point out that "visual" properties do not determine content or one's knowledge of it (*Blue Book*, p. 39 and passim).[12]

How could a skeptical conclusion about our knowledge of our own thoughts come to seem unavoidable? By way of setting up the case for the deflationary analysis, Boghossian describes the options for the understanding of self-knowledge as exhausted by three possibilities: such knowledge is either based on inference, or by a kind of looking, or else it is based on nothing (p. 5). With respect to knowledge of the content of one's thought, and much other self-knowledge, it seems

[12] In brief, I think Peacocke, for instance, is right in claiming that "it is a datum that we do know the full, ordinary, externally individuated intentional content of our own thoughts, and of other people's utterances, without reliance on inferences from, or presuppositions about, something weaker, which is all, in some alleged stricter sense, we would be aware of on the internalist introspectionist's view" (1998, p. 79). The reference to "something weaker," i.e., something with intrinsic recognizable features of its own, is crucial to the perceptual model.

that the first possibility cannot be right. If introspective awareness is anything at all—that is, anything distinct from the knowledge of the mental life of others—then it seems it must be something different from any knowledge based on inference. As to the second possibility, examples such as that of observing the coins are taken to show that, if externalism is true, we cannot know our thoughts by inward "looking" either.[13] That leaves the option of seeing self-knowledge as "based on nothing." This, then, is taken to mean that so-called self-knowledge cannot in fact be seen as a "cognitive achievement" of any kind, and cannot sustain what he calls a "substantial epistemology." To reject a substantial epistemology for self-knowledge is to reject any form of the idea that it involves the awareness of some independently obtaining state of affairs. Boghossian briefly discusses some examples of what he means by "insubstantial" knowledge, such as the indexically grounded judgment that "I am here now," all of which examples share the feature that the appearance of knowledge is grounded purely logically (or transcendentally), and hence that the denial of any such statement would involve some kind of immediate incoherence. What this means in fact is that avoiding the "insubstantial" conclusion is even more urgent than the paper suggests. For it would not just be disappointing or deflationary if self-knowledge were to turn out to be insubstantial in this sense. (Philosophers, after all, are supposed to be hardened to such disappointments.) Rather, such a conclusion would just clash with the kind of statement being made in an expression of self-knowledge, for there is generally *no* logical incoherence in the denial of a first-person statement of some attitude. So a statement such as "I believe I was born in Minnesota," which has the appearance of an expression of self-knowledge, cannot be cognitively insubstantial in the logical or transcendental sense described. The claim to knowledge of one's own belief here could be doubted or denied without the incoherence that

[13] Naturally, one might balk at this conclusion, too. For instance, if a person's identity is constituted by certain causal-historical facts about his birth, etc., does that mean I cannot recognize someone by looking at him?

would follow upon denying the truth of "I am here now," and hence cannot be "insubstantial" in that sense.

However, the argument combines two quite different senses of "cognitive achievement" in order to raise the skeptical possibility. If we consider what makes the above examples cases of "insubstantial" judgments, it would seem that by contrast a genuine cognitive achievement requires that its truth conditions be in some way independent of the making of the judgment (as, arguably, they are *not* in a statement like "I am here now"). This is a general form of cognitivity that any account of introspection as a source of knowledge would seek to preserve. However, the beginning of this section of the paper assumes a definition of "cognitive achievement" as knowledge of a contingent proposition that involves either observation or inference from some observation (p. 17). And that is quite a different matter. For any definition of this latter kind clearly makes it impossible to conceive of self-knowledge as both "substantial" but not conforming to the picture of "inner observation." Thus, the argument from the three options offered for understanding the status of self-knowledge assumes the appropriateness of that picture and relies on it to raise its skeptical challenge.

There's another way of looking at the "insubstantial" conclusion, of course. If a perceptual model is *not* tacitly assumed and we take self-knowledge to be "insubstantial" only in this second, stipulated sense (i.e., "contingent knowledge not based on observation or inference"), then it may well be a conclusion to be welcomed and not avoided. By itself the conclusion would pose no skeptical threat, for all it means is that introspective awareness is "immediate," in the sense of noninferential, and in addition is not to be construed as a form of perception. One may of course stipulate such a sense of "insubstantial," but then we must recognize that nothing in it is per se incompatible with the fully cognitive status of the judgments in question.

There may, of course, still be difficulty in conceiving of the possibility of apparent judgments that are "based on nothing" in the above sense, but which still represent a genuine cognitive achievement of some kind. And that difficulty, combined with the assumption of the

perceptual model, presumably contributes to the conflation of the two senses of "insubstantial." But there are in fact several aspects of one's relation to oneself as an agent which have been plausibly seen as involving awareness that is not based either on behavioral inference or any perceptual presentation.[14] A person is commonly aware of his own basic movements and bodily position without having to observe anything, internally or externally. There need not be any characteristic internal sensation present, and even when there is, it is not that on which the person bases his judgment that, for example, his knee is bent. I do not mean simply to take the idea for granted, and the case of action will receive further attention in Chapter Four, but I do take it to show that the considerations discussed above do not in any way force us to some deflationary account of first-person reports. At the very least, the burden of proof would be on someone who claims that we must adopt a perceptual model if we are to see self-knowledge as involving a cognitive achievement at all. In this respect we might compare the awareness one has of one's bodily position with a case like that of judging what time it is.[15] These are judgments of contingent fact which can, of course, be made on the basis of observation and evidence of various kinds (looking at a clock, or the position of the sun), but in a central range of cases they may involve no such observation at all. You perhaps shut your eyes, consider the question, and deliver an answer. And in cases like these, such judgments share just the features that Boghossian later cites as the earmarks of genuine cognitive achievement, and which distinguish them from the earlier cases of "insubstantial" or self-verifying judgments (p. 19). That is, these judgments are subject to the direction of one's attention, and the accuracy of one's judgment is normally dependent on the exercise of such atten-

[14] For an initiating discussion, see Anscombe's 1957 book on intention, in particular §§8 and 28 on 'nonobservational knowledge'. For some more recent discussion of the awareness we have of our intentional bodily movements, see chapter 5 of Wilson (1989, esp. pp. 121–24), and Peacocke (1992, pp. 90–96).

[15] See Wittgenstein's extended discussion of this case in *Philosophical Investigations* (1956, §607).

tion. Some people may in general be better at such "detection" than others, and there is naturally no trouble here in conceiving of the possibility of error, or how it might be corrected. But, nonetheless, neither the judgment that one is sitting down nor the judgment that it must by now be nearly noon need be based on any quasi-perceptual presentation. One may choose to describe this situation as one in which such judgments are "based on nothing," but that would not be equivalent to denying that they are cognitive achievements.

These considerations do not dispose of the perceptual model, of course, but they should suggest that this model is an optional one and is not forced on us simply by the guiding assumption that first-person judgments are genuinely expressive of a kind of awareness. Tacit assumption of the perceptual model plays a role in encouraging some kind of deflationary account of first-person discourse, both for reasons of general hostility to the "inner eye," and, as we have just seen, for more sophisticated reasons having to do with the theory of mental content.[16] The perceptual model has problems quite independently of these concerns, however, and it is crucial that we keep the distinction clear and do not take the substantiality of self-knowledge itself to be identified with a particular intuitive model of it.

1.3 CONSTITUTIVE RELATIONS AND DETECTION

We are pursuing an understanding of self-knowledge that would make sense of both success and failure in introspection; that is, account for a person's introspective attempt to get something right, allow for the possibility of error and ignorance, and thus accommodate some independence of awareness and the object of awareness. The comparison cases just mentioned concerning judgments of time and bodily position

[16] Again, the dialectical situation of Boghossian's own paper suggests that it is Externalism about content that is his real target, rather than the reality of self-knowledge itself.

might be thought to provide models of judgments of substantial contingent matters which need not be perceptually based. Or, if not providing models of introspective access, they might at least calm fears that such a category of judgment is a conceptual impossibility. But, at the same time, such cases suggest a purely contingent connection between the obtaining of the states of affairs in question, and the fact that we are so constituted so as to be reliable detectors of them (we know not how). The situation with respect to awareness of one's mental life has usually seemed to be something quite different. It doesn't seem that we just happen to be wired up in such a way as to be reliable reporters of our pains, intentions, or feelings of anger; and it is difficult to conceive of being a proper subject of such states but only being able to become aware of them in a third-person way. Rather, it has seemed to many philosophers to be central to our very concepts of these states that the person's own reports of them should be both "immediate" in the sense defined, and enjoy a certain authority over the reports of others; whereas by contrast it is no part of our concept of time that we should be particularly good detectors there, let alone good detectors who don't need to rely on perceptual evidence (the case of bodily position raises different questions). And the claim that, say, it is essential to a class of mental states to be available to the person introspectively has been taken as a further part of the case against viewing it as a form of perception, since in the case of external perception there is only a contingent connection between the existence of the objects and any awareness of them.[17]

There is, at any rate, a strong suspicion of a *conceptual* requirement lying at the bottom of first-person authority, and the a priori nature of such a requirement has suggested to many philosophers a different set of reasons for thinking that the "authority" in question cannot be a genuine or substantial one. On many such views, the appearance of reliable discovery of one's own mental states is in fact merely the shadow cast by certain features of our linguistic practices. Since the

[17] A point emphasized by Shoemaker in recent papers.

following chapters will be developing an account of self-other asymmetries that takes them to be essential to the nature of persons generally, I want first to investigate the prior question of whether admitting some conceptual basis to first-person authority undermines the assumption of first-person reports as involving genuine cognitive achievements.

Earlier we saw how the tacit assumption that introspection must be perceptual if it is substantial at all can lead one, perhaps reluctantly, to the conclusion that some kind of deflationary account is inevitable. In recent work of Crispin Wright, we encounter a symmetrical movement that *starts* from a principled rejection of any perceptual model and moves from there to the development of an account of first-person discourse explicitly designed to be deflationary. 'Deflation' in this context means that either first-person psychological discourse is interpreted as not *reportive* at all, or the 'authority' of such statements is seen as having some wholly *non*cognitive basis. This is thought to be a consequence of an account that avoids the perceptual model by stressing a set of a priori conceptual connections between mentality and first-person authority. For my purposes, then, it is crucial that we keep separate the questions of conceptual dependence and the question of the substantiality of self-knowledge. It is equally important to see how, even on deflationism's own terms, the particular features of the first-person position are simply left out of such an analysis.

In a series of papers, Wright has developed an account of first-person psychological discourse that is designed to account for the privilege normally accorded such statements, while avoiding the implication that such privilege expresses recognition of any properly epistemic virtue of the first-person position. On this account, our concepts of various mental states make first-person judgments of them "extension determining," in the sense that a person's best opinion about his intention (Wright's example) does not detect or "track" that state, but rather, for a priori reasons, determines its identity. In this way, the conceptual connection between mental state and first-person judgment is very tight, for the latter determines the former. Wright be-

gins working out this account after canvassing various Wittgensteinian objections to conceiving of one's knowledge of one's own intentions (and other mental states) as involving a kind of inner perception.

> So far as I can see, there is only one possible broad direction for such an explanation to take. The authority which our self-ascriptions of meaning, intention, and decision assume is not based on any kind of cognitive advantage, expertise or achievement. Rather it is, as it were, a *concession*, unofficially granted to anyone whom one takes seriously as a rational subject. It is, so to speak, such a subject's right to declare what he intends, what he intended, and what satisfies his intentions; and his possession of this right consists in the conferral upon such declarations, other things being equal, of a *constitutive* rather than descriptive role.
>
> (1986, p. 401)

The constitutive role of first-person judgments is made out in terms of the distinction between extension-determining and extension-reflecting concepts. The former notion is explicated by comparison with the familiar analysis of color as a secondary quality. That notion can be expressed in terms of sets of biconditionals, whose truth is knowable a priori and which fix the meaning of the concepts in question. So, for a color concept like 'red', for instance, we specify such things as a set of normal perceivers and a set of conditions favorable for the making of color judgments, and then arrive at a biconditional of the form:

> In conditions C: *X* is red iff *X* is judged to be red by normal perceivers.

The details of such conceptual analyses have been the subject of much contemporary work on the proper understanding of realism and related issues, and need not concern us here. I am concerned with the application Wright makes to the case of self-knowledge and the conclusions he draws from it. The application he makes is quite straightforward. We can account for the acceptance of first-person authority with respect to intentions by seeing our concept of intention as constrained

by a priori biconditionals of the following form. Assuming a normal psychological subject, and given the appropriate conditions of attention, mastery of the relevant concepts, and so on,

S has the intention to **Φ** iff *S* judges that he intends to **Φ**.

Wright takes such biconditionals to express our genuine conceptual commitments with respect to intentions and other psychological states, and he sees the a priori status of the biconditional as incompatible with seeing first-person judgments of intention as extension-reflecting, rather than extension-determining. If such judgments of intention were extension-reflecting, that is, involved the genuine detection of some independent state of affairs, then any a priori declaration about their reliability would be unwarranted (1989, p. 253). Thus, the claim of extension-determination is offered as the best explanation for our a priori commitment to biconditionals of the form described.

This, in abbreviated form, is Wright's case for conceiving of first-person authority in terms of social concessions rather than in terms of cognitive advantage. Once again, but now in a different way, self-knowledge is said to fail of a "substantial epistemology." The cogency of this deflationary analysis clearly depends on the avoidance of trivializing specifications of the C-conditions, so that they do not simply build in "whatever conditions are required for the accurate detection of one's intentions." This problem gets a good deal of attention in Wright and in some of the subsequent literature (cf., especially, Holton 1993). The "insubstantial" conclusion depends equally on the case for the a priori status of the biconditional, for if instead it had the status of a good empirical generalization, then obviously the case against genuine first-person *detection* of intention would not be made. In addition, there is a surprising transition from claims about judgment-dependence of the sort represented in the biconditionals to the claims from the quoted passage against the appearance of cognitive achievement and in favor of an analysis in terms of social concessions. The inference is surprising because the original idea of the extension of certain concepts being partially determined by the judgments of appropriately placed appliers of those concepts begins life in its application to the case of secondary

qualities, such as colors. But it would certainly not follow from any such analysis of color concepts that particular judgments of the color of something were not expressive of a cognitive (indeed, perceptual) achievement of some sort, and were instead a matter of some kind of social concession. So, even if the relevant biconditionals for intention could be specified nontrivially and their a priori status were secured, this would not serve to show that first-person authority was not based on some kind of genuine cognitive advantage. Nor would it even serve the purpose of ruling out a perceptual model of introspection, as the color analogy shows.

For our purposes, however, the chief weakness of any analysis of this sort is how little it ends up illuminating any of the familiar *asymmetries* between first- and third-person psychological discourse. After all, response-dependence of some form or another is a feature of a great variety of concepts.[18] Nothing in the analysis itself given here explains why there should be any difference at all in the application conditions of psychological concepts in first-person and third-person contexts. In fact, as far as the analysis goes, there may not even *be* any such difference. For we could specify a similar set of biconditionals as governing the application of psychological concepts to *others*. That is, we could specify C-conditions, competent ascribers, conceptual capacities, and so forth, in such a way as to make it an a priori matter that such ascriptions have a strong prima facie claim to truth. Indeed, on a common understanding of the sort of "interpretation theory" associated with Donald Davidson and Daniel Dennett, this is precisely how things stand with respect to (commonsense) psychological attributions generally.[19] The account of extension-determining concepts was pre-

[18] The idea of "response-dispositionality" as a constitutive feature of certain fundamental concepts has a wide literature by now. See, for example, recent papers by Johnston (esp. 1991).

[19] One need not go all the way in the direction of Dennett's instrumentalism to see commonsense psychological concepts as part of the class of response-dependent concepts.

See Moran (1994) for further discussion of Interpretation Theory in the philosophy of mind and the question of whether psychological discourse in general is to be understood as the application of an explanatory theory (the "theory theory" of mental terms).

sented as an account of first-person authority, and yet it does not help us to understand why it is that *first*-person ascriptions should have any *special* claim to truth. Nor does it account for why such ascriptions are routinely made without reliance on evidence, unlike their third-person counterparts. For all the biconditionals tell us, it could be that first-person ascriptions were only made on the basis of examining the behavioral evidence (i.e., one's own), but our convention dictated that we always privilege the person's own reading of that evidence as being the best possible one. Finally, any adequate analysis of the first-person would have eventually to get beyond the picture of "privilege" and "concessions" and say something about how the presumption of first-person authority expresses an ordinary rational *demand* quite as much as it reflects any deference to the person's best opinion about his own state of mind. We do not only allow his statement to stand without the benefit of evidence, we also expect and sometimes insist that he take himself to be in a position to *speak for* his feelings and convictions, and not simply offer his best opinion about them. ("Do you intend to pay the money back?" "As far as I can tell, yes.") And it is part of this same demand that not only do we not expect the person to need to base his statement on evidence, but we may regard his deferring to the behavioral evidence as a form of evasion, or else as suggesting that the state of mind he's reporting on cannot be a fully rational one.

This normative expectation, and its relation to the rationality of the beliefs in question, certainly lends some support to the suggestion that the first-person accessibility of beliefs is not a *merely* empirical matter, an extra capacity for awareness of a certain class of facts we happen to have and whose absence would leave the psychological facts in question unaffected. This suggestion is developed at greater length in Chapters Three and Four. But the nonempirical or conceptual aspect of the phenomenon does not support either a conventionalist reconstruction of first-person authority, or a deflationary analysis of the claims of self-knowledge. That psychological attributions gener-

ally presume a background of rational intelligibility is an assumption well entrenched in contemporary philosophy of mind and is widely accepted by both realists as well as instrumentalists about psychological phenomena. Introspective awareness could be perfectly substantial, even if the assumption that a person's mental life is accessible to him in this special way has a basis that is partly normative and conceptual.

1.4 "CONSCIOUS BELIEF": LOCATING THE FIRST-PERSON

It ought to seem surprising that accounts of first-person authority should fail to characterize or account for a distinctively first-person relation. I want to suggest that part of the reason for this is a concentration on the parallel (as well as the disanalogy) with the situation of making judgments about the external world. This encourages a purely theoretical model of the situation of introspection, a concentration on questions of belief and judgment as applied to some static realm of mental facts. Inadequate attention is given to the person as epistemic agent, and hence to the mutual interaction between mental life and the first-person awareness of it.[20] We saw this theoretical model at work in the case of Boghossian's tacit assumption of the Perceptual Model of introspection (which may be seen as one graphic version of what I am calling the general theoretical model), and now in Wright's conceptual analysis, which describes the abstract conditions for making a set of judgments (i.e., psychological attributions) but which leaves out of account why the presumption of truth should be any different, or differently based, in the first-person case. Neither analysis connects the privileging of first-person judgments with any of the wider asymmetries, including the distinctive *manner* in which they are made. But the prob-

[20] A major exception here is the recent work of Tyler Burge on the first person. See especially "Our Entitlement to Self-Knowledge" (1996) and "Reason and the First Person" (1998).

lem of self-knowledge is not set by the fact that first-person reports are especially good or reliable, but primarily by the fact that they involve a distinctive mode of awareness, and that self-consciousness has specific *consequences* for the object of consciousness.

What I mean by the concentration on the theoretical has only been sketchily indicated so far and will become clearer by consideration of a final representative account of self-consciousness that most explicitly declares its allegiance to this picture. It is a commonplace in discussions of self-consciousness to conceive of the target notion in terms of second-order states, but D. H. Mellor's "Conscious Belief" seeks to build this assumption into a complete account of the phenomenon in question. Mellor begins with what he calls an "action theory of belief," a view that needn't be disputed here, since it can be understood to stand for the basic assumption that the notion of belief (and related states) is tied to its role in the explanation of behavior. What this assumption most directly opposes itself to is the idea that belief and the like are intrinsically identifiable phenomenal states, a view that has few adherents today. It will follow from an action theory of belief that at any moment a person may be described in terms of a host of dispositional states, tacit beliefs, assumptions taken for granted, as well as explicit beliefs and desires, which together contribute to his immediate thinking and behavior. Not all of this will (or even could) be conscious at any time; some of it never will be. Hence, we need a term to describe "the new state of mind I come into when a belief of mine becomes a conscious one" (p. 88), and Mellor settles on the term 'assent'. What he describes as his main thesis, then, is "that assenting to a proposition is believing that one believes it" (p. 90).[21]

Part of what drives the analysis of conscious belief in this direction is simply the fact that "being conscious of" is undeniably a cognitive relation of sorts. It is a way of knowing something, or having it available for further thought. But, nonetheless, with respect to the aware-

[21] His secondary thesis is that linguistic action—speech and writing—requires second-order beliefs.

ness of mental phenomena, the case of second-order beliefs is too broad to capture either the particular character of conscious awareness or the specifically first-person character of conscious belief. As to the first point, consider the ordinary phenomena of either tacit or unconscious beliefs. A tacit belief may be something the person takes for granted but has never reflected on explicitly. Contrary to what Mellor suggests (p. 93), such a state is not simply a disposition to assent to the proposition in question (a disposition which, since it doesn't require the cooperation of a desire in order to assert itself, Mellor would place outside the class of genuine beliefs). Rather, tacit beliefs may interact with relevant desires to produce action in much the same way as do explicit beliefs. Problem solving (a desire-guided action) often requires making explicit some tacit assumptions that stood in the way of a solution. The person's action in pursuing a particular misguided line of thought is explained by reference to his maintaining the faulty assumption, and explanation of his pursuit of that false trail requires ascribing this assumption to him. And, far from the tacit belief's being a mere disposition to *assent*, its coming to consciousness in such cases is accompanied by immediate *dis*sent from the proposition one had been taking for granted.[22] It was thus a belief maintained only on condition of its not being a conscious one.

Now, if it's agreed that first-order beliefs can be either unconscious or tacit beliefs, then there's no reason why the same cannot be true of second-order beliefs themselves. Thus, for instance, a person may take it for granted that his friend, like most people, believes that dead men tell no tales, even if this thought about his friend has never crossed his mind. And, of course, he may equally well take it for granted that he himself does not differ from his friend in this respect, again without the thought ever occurring to him. In both cases, then, he has a belief about someone's belief without its ever occurring to consciousness. And in the second case it's not just a belief about someone's belief, but

[22] See Stalnaker (1984) for a discussion of such cases and their implications for the nature of belief and belief-attribution. See especially p. 69 and passim.

it is a *first-person* second-order belief, which, for all that, is still not a conscious belief. Similar considerations apply to the case of beliefs that are not tacit but are unconscious for perhaps more psychologically interesting reasons.

It would seem, then, that the particular features of conscious awareness of belief (what Mellor calls 'assent') cannot be described merely in terms of second-order beliefs; not when my belief concerns someone else's mental state, nor even when it concerns my own. However, more important than failing to capture the idea of explicit awareness in the account of conscious belief is the absence of any sense of what sort of difference is made by the distinctively *first-person* awareness of one's belief. We may think of the problem in the following way. We saw that I can have a second-order belief whose object is someone *else's* first-order belief, without that involving an episode of explicit awareness at all. What I now want to point out is that an analysis like Mellor's is not rescued by refining it so as simply to require explicit, episodic awareness of belief. The particular first-person character of conscious belief would still be missing. For, of course, I could be explicitly, consciously reflecting on my friend's belief about life on Mars without that making it a conscious belief of anyone's. Nor need it make any essential difference if it were my own belief that I was consciously reflecting on, if I attribute it to myself under a name or description I don't recognize. Nor even if I knowingly attribute the belief to the person I recognize as myself, using the first-person pronoun, but, say, ascribe it to myself only on the basis of reading a letter I wrote last night (where I have reason to believe I still retain this belief, even though I can't now remember why).

What any of this would leave out is the fact that to call something a conscious belief says something about the *character* of the belief in question. It is not simply to say that the person stands in some relation of awareness to this belief. If someone is looking at a tree, referring to it as an "observed tree" would not express anything about its qualities as a tree, and similarly with the unspecified awareness of someone's belief. By contrast, a conscious belief enters into different relations

with the rest of one's mental economy and thereby alters its character. We speak of the 'consciousness' in 'conscious belief' as something that informs and qualifies the belief in question, and not just as specifying a theoretical relation in which I stand to this mental state. If it were simply a special immediate theoretical relation I have to this belief, then there would be no reason in principle why another person could not bear this same relation to my belief. But in such a case *my* belief would not thereby acquire the attributes we have in mind when we apply the term 'conscious' as a characterization of the belief itself. (Think of an unconscious attitude of resentment. If I become aware of it only because I fully believe the interpretation given by my analyst, the attitude does not thereby become a conscious one. There is still work to be done.)

We apply the term 'conscious' to the belief itself for reasons related to why we may apply this term to certain activities of the person, where this qualifies the activity in ways that do not obtain with respect to anyone else's awareness of it. To play the piano either attentively, or unreflectively, or deliberately to annoy someone, makes a difference to the quality of the playing. In cases like these, the cognitive terms used denote adverbial modifications of the activity itself. Similarly, it is only with respect to one's *own* activities that 'consciousness' has such an adverbial function; so that, for instance, sleepwalking, walking normally and unreflectively, and walking with conscious deliberateness are all distinct kinds of activity. In this last case, the person's consciousness of his activity is not something that stands outside it observing, but infuses and informs it, making a describable difference in the kind of activity it is. In a related manner, when my assumption that the person just referred to as "Sue" is not a boy becomes a conscious belief, this change makes a difference to its relations with my other beliefs, and to my confidence in it. My whole relation to this assumption is now different, and the belief itself no longer has the secure, taken-for-granted quality it had before. And in both cases the empirical, qualitative differences made by an activity or an attitude being a conscious one are bound up with the differences in the person's autonomy and

responsibility. Just as unconsciously standing on someone's foot in a crowd is different from doing so in full awareness, so for an attitude to be a conscious one makes a fundamental difference to the person's relation to it, in addition to the bare fact of awareness and whatever empirical difference that awareness may make to the character of the attitude.

The special features of first-person awareness cannot be understood by thinking of it purely in terms of epistemic access (whether quasi-perceptual or not) to a special realm to which only one person has entry. Rather, we must think of it in terms of the special responsibilities the person has in virtue of the mental life in question being *his own*. In much the same way that his actions cannot be for him just part of the passing show, so his beliefs and other attitudes must be seen by him as expressive of his various and evolving relations to his environment, and not as a mere succession of representations (to which, for some reason, he is the only witness). And in both the case of actions and attitudes, self-consciousness makes a difference to what the person's responsibilities and capacities are, with respect to his involvement in their development. It is modeling self-consciousness on the theoretical awareness of objects that obscures the specifically first-person character of the phenomenon, whether or not this theoretical perspective takes the specific form of the perceptual model of introspection.

What we have so far characterized as the specifically first-person manner of awareness that qualifies a belief or other attitude as a conscious one is an awareness that is immediate, nonobservational, and involves reference to oneself through use of the pronoun 'I', rather than by means of some mediating description under which the person might fail to recognize himself.[23] But we will need a fuller characterization than this to account for the special features of first-person awareness. For it would still be possible for a person to have immediate awareness of an attitude of his that conformed to the above conditions but was still essentially a kind of outsider's perspective on his attitude.

[23] There is a wide literature here. See especially Perry (1977 and 1979) and Shoemaker (1968).

That is, the conditions so far specified could still apply to a case where, say, I had immediate awareness of my attitude (perhaps in the way one has immediate awareness of the disposition of one's limbs), but where the attitude was one of which I could make no sense, or whose reasons were opaque to me. The attitude could be one which I couldn't link up with other attitudes of mine, and which persisted unaltered by and in isolation from both my own criticism of it and my explicit reflection on the object that the attitude is supposedly directed upon. For such reasons, it might well be an attitude that I would not allow to play any explicit role in my deliberations about the object in question. Thus, if the attitude in question is a belief, it would then be a belief I was conscious *of*, but it would not have any of the first-person character that is indicated by referring to something as a conscious belief. Even if immediate, my consciousness of it would be just as external to it as the immediate awareness someone might happen to have of some internal bodily process. A person who only had awareness of his mental states that was immediate, but alienated, in this way could not be said to "know his own mind" in the sense we take for granted in ordinary life. If such a person would lack ordinary self-knowledge, then something is missing from our original characterization of the target notion.

A more complete characterization of the first-person perspective will require bringing the agent more explicitly into the picture, and doing so will involve taking the discussion into a range of issues concerning the agent's perspective of deliberation and self-interpretation that have not been at the center of recent discussions of self-knowledge. One thing that is unsatisfying about any perceptual model of self-consciousness is that perception is a relation that, in principle, should be possible with respect to a whole range of phenomena of a certain type. On such a model, then, there would seem to be no deep reason why one couldn't bear this quasi-perceptual relation to the mental life of another person as well as oneself. From this perspective, the restriction of this mode of access to the *first*-person is unmotivated and not essential to

it.[24] Not every form of self-knowledge has a claim to "inalienability," that is, can be shown to be a form of apprehension that is essentially and exclusively first-personal in its reference, but the forms we are concerned with here do raise the question of why this sort of apprehension should be restricted at all in its scope. What I mean by restriction of scope is primarily two things. First, there is the question of why the kind of apprehension we have characterized so far should be reserved for awareness of only one's own mental life. And this requires doing better justice to the "reflexive" aspect of first-person awareness than we have done so far, including the relation of self-knowledge to some of the special features of self-interpretation that have attracted attention elsewhere in philosophy. Second, there is the restriction in scope, not to a particular person, but to a particular *class* of facts about oneself, that characterizes this form of awareness. That is, whatever the "authority" of the first-person is, we will want to understand better why the person is assumed to speak with such authority (when he is) with respect to facts about his *psychological* life, and not with respect to facts about any other "inner" processes (physiological, neurological, etc.).[25] A distinguishing fact about "intentionally characterized" phe-

[24] For some philosophers, that is precisely the point, or one of the points, of the defense of the perceptual model. Armstrong (1984), for instance, explicitly defends the possibility of "introspective access" to the mind of another person (pp. 113–16). In a more recent paper, Crispin Wright insists, as part of his rejection of the perceptual model, on the "inalienable" character of any self-knowledge characterized by first-person authority: "The kind of authority I have over the avowable aspects of my mental life is not transferrable to others: there is no contingency—or, none of which we have any remotely satisfactory concept—whose suspension would put other ordinary people in a position to avow away on my behalf, as it were" (1998, p. 24). A footnote to this sentence denies that "we have any satisfactory concept of what it would be to be in touch with others' mental states *telepathically*." My agreement here will already be clear, but development of this aspect of first-person authority will be taken up more fully in Chapters Three and Four.

[25] Gareth Evans and others have argued that other sorts of facts, e.g., facts about one's spatial location or bodily nature, may be known "immediately," and my formulation is meant to be neutral on this question. It is less clear whether any asymmetries of authority are thought to obtain with respect to such facts.

nomena generally (not only states of mind, but actions, practices and institutions, including linguistic ones) is that they admit of a distinction between inside and outside perspectives, the conception of them from the point of view of agents or participants as contrasted with the various possible descriptions in some more purely naturalistic or extensional idiom. The phenomena of self-knowledge participate in this duality of perspective, and it is only under some descriptions and not others that the person's own description of his state is accorded any kind of privilege. If self-knowledge is indeed a form of *knowledge*, then it will be constituted by the person having thoughts about his state, and even more basically, by his *conceiving* of himself and his state in certain ways. So, prior to understanding 'first-person authority' in terms of superiority of access, we will want to understand why the person's *conception* of his state of mind has been thought, in certain contexts, to play a privileged role in making the state what it is. The following chapter takes up various senses of 'privilege' that have been thought to characterize a person's own conception of his thought and action, particularly as these are seen as having some "self-constituting" role. Exploring the bases and limitations of such a role brings the agent's perspective more squarely into the discussion of self-knowledge by way of relating the authority of the first-person to the role of the deliberator in determining his state of mind. My hope is to provide the terms for a more detailed and realistic picture of the ordinary failures as well as successes of self-knowledge, and why some of the characteristic failures should matter in any special way to the health of the person.