8

Self-Knowledge, 'Transparency', and the Forms of Activity

Richard Moran

1. Introduction

Traditionally, the discussion of self-knowledge and self-consciousness in philosophy has given it a place central to the understanding of rationality and agency themselves, and in doing so it treats the ability to know one's own mind as something more than a useful capacity we enjoy as human beings. In a contemporary context, this is a thought more often in the background of philosophical discussion than something defended explicitly. It is nonetheless an assumption recognizable enough to be found debatable, or difficult to make sense of in the context of contemporary understandings of the nature of mentality and agency themselves. This chapter begins by setting out in very general terms some considerations that would link self-knowledge to a certain form of agency, and then considers two recent studies that seek in their different ways to show that a certain form of account of self-knowledge (involving appeal to the "transparency" of belief) must be divorced from any appeal to rational agency. I will be arguing that in both cases the account that emerges from this divorce ends up with a kind of agency in the picture after all, only of the wrong kind, so that the ordinary exercise of coming to know what one believes

requires the person to exert a kind of external control over her own attitudes. In seeking to develop the relevant notion of rational agency in contrast to this, I cannot claim to be doing more than pointing to the place where I think we need some such notion. In the course of doing so, I try to characterize the sense of 'activity' or 'agency' that is relevant to a central class of cases of self-knowledge, and distinguish this sense of activity from the sense of activity indicating a process of production, or acting upon oneself so as to produce a belief.

In thinking about self-knowledge and rational agency there are two broad directions from which we may begin to ask how they may be related to each other. We may ask, first of all, how self-knowledge matters to agency itself, that is, whether the specifically human forms of *rational agency* can be understood apart from the capacity for self-knowledge of the mental life that is expressed in that agency. Is our capacity to act for reasons, to be self-guided in that sense, dependent on our ability to know our mental life 'immediately'? Can the ordinary ability to respond to reasons in one's thinking, to consider reasons for and against some belief and respond accordingly, be understood apart from our capacity for immediate self-knowledge? And would the absence of the ordinary capacity for self-knowledge make no essential difference to our rational agency?[1] And from the other direction we can ask how rational agency itself may matter to the understanding of self-knowledge; that is, whether the ordinary capacity to know what one thinks about something is part of the same capacity to determine one's thought about that thing. Is our ability to know what we believe 'immediately', and with a kind of authority not shared by what we say about the beliefs of others, tied to the fact that our beliefs and other attitudes are expressions of our rational agency, and is there a notion of responsibility applying to a person's relation to her attitudes that is related to the capacity for first-personal knowledge of them?

One recent way of relating both sets of questions begins with the example of belief, and appeals to a notion of 'transparency' between a question about one's belief and a corresponding question about the object of one's belief. Thus, it has seemed to several philosophers that a distinctive feature of first-person discourse is that a person can answer a question about her own belief by addressing herself to the corresponding question about the topic of that very belief.[2] Hence, if asked do I, R.M., think it will rain today, I can answer this question by giving my answer

[1] Recent philosophical work has drawn connections between self-knowledge and agency in a variety of different ways. Here I will just mention Burge 1998, Bilgrami 2006, and O'Brien 2007. Particularly helpful to me in thinking about the issues of this paper has been Matthew Boyle's paper 'On 'Making up your Mind and the Activity of Reason' (forthcoming).

[2] See Evans 1982, Edgley 1969, and Moran 2001. More recently Byrne 2005 and Shah and Velleman 2005 appeal to a notion of 'transparency' for belief, but offer very different accounts of it.

to the corresponding question about the rain, and not by inquiring in to the state of mind of a particular person. The fact that the answer is given by a particular person, the very person whose state of mind the first question directed itself to, must surely be part of the answer to how it is possible or legitimate to answer the question about one's belief by reference to the question about rain. For 'transparency' of this sort surely does not apply to a question I may ask about the beliefs of another person. If the topic of my question is the beliefs of some other person, then my efforts to answer that question must address themselves to the facts concerning the state of mind of that particular person. That's how it is, after all, with my efforts to answer questions about other topics. In seeking to answer a question about the inflation rate in China, I must direct my attention to China. Why, then, if the topic is some fact (attitudes of a certain kind) about oneself is it legitimate to answer the question in a way that seems to neglect the fact that it is about a particular person, and instead treat it as a question concerning the topic of the attitude itself (e.g., the weather)? Again, the fact that in such a case the answer is delivered by the very person whose state of mind is in question must be central to accounting for this, for if transparency is ever legitimate it must represent a systematic difference between relations to oneself and relations to others. The identity of the person whose state of mind is inquired into and the person answering the question must matter here.[3]

The form of account I give appeals to a form of agency that is part of a person's being a creature with beliefs, and I claim that the transparency, which various philosophers have found attractive, cannot be accounted for without appeal to this agency. It would not, in general, make sense to answer a question about my state of mind (e.g., my belief about the weather) by attending to a logically independent matter (the weather itself) unless it were legitimate for me to see myself as playing a role in the determination of what I believe generally, not in the sense that beliefs typically owe their existence to acts of deliberation but that the responsiveness to reasons that belongs to beliefs is an expression of the person's rational agency. However, while this form of account takes the topics of self-knowledge and agency to be closely related, the agency in question does not involve any kind of voluntarism about belief, and indeed the form of rational agency I have in mind has as a consequence that such voluntarism is false. The sense in which I see belief and other attitudes as forms of activity is deeply related to the fact that they are not matters of choice for the person, and hence the agency involved here is not that which is exercised when, say, a person chooses to raise her arm and then does so.

[3] Elsewhere, I have presented an account of self-knowledge that seeks to vindicate the applicability of transparency to the first-person case, and with it the claim that in delivering an answer in this way, the person is indeed speaking from knowledge (Moran 2001).

2. Knowing What and Knowing Why

We can take the case of belief as representative of the attitudes generally, though it will be important to recognize that there will be differences between, for instance, believing, wanting, fearing, intending, hoping, and caring about. What they share as attitudes, however, is their involvement in forms of normative assessment, such as that of a belief's being justified, or a fear's being unwarranted, or an activity's being worth caring about. In this they differ from brute sensations, which I take to be aspects of our passive or receptive nature.[4] Part of what is meant by this is that different forms of the question 'Why?' will apply to items in these two categories, and the relevance of asking the question of the person herself will also be different. There may be no special reason to ask me why I am experiencing pain in my lower back, although I might know something about it. And the answer to that question 'why?' will refer to such things as a previous injury, bad work habits, or compression on a disk. That is, the answers will contribute to the explanation of the coming to be of the sensation I am experiencing, but do not seek to say anything about the apparent point, or the good, or the intelligibility of the state I am in. Naturally, such causal versions of the question 'why?' will also apply to items in the other category as well (e.g., one's beliefs, hopes, desires, and intentions), but it is internal to them that a different form of the question 'why?' also applies to them. Hence, we can ask why someone believes it will rain when we are asking for that person's reasons for the belief, and we can ask why someone wants a saucer of mud when we want to know what could seem good or worth having in such a thing. This form of the question 'why?' thus seeks a certain normative characterization of the attitude itself, seeing it as reasonable or not, worthwhile or not. However, just as with the question seeking the origins of some condition, the answer to the question 'why?' is meant to tell us something relevant not just about the character of the attitude itself, but something relevant to the fact of the person's having that attitude. That is, the way in which the attitude is found reasonable or intelligible by the person is assumed to be relevant to the question *why she has that attitude*, how it came to be part of her mental life, or what maintains it there. So this is something that this question shares with the kind of question 'why?' that applies to the person's sensations, in that in both cases we are inquiring into how something comes to be or maintains itself. The difference is that in the case of the question 'why?' as applied to the person's beliefs and intentions, the question of how it comes to be is tied to this normative

[4] Ernest Sosa (this volume) makes a similar point with respect to the difference between experiences and "seemings": "The seeming is rationally based; the experience is not. The seeming manifests the subject's rational agency; the experience does not, any more than does a pain."

question involving notions like the reasonable and the worthwhile. The second broad type of 'why?' question assumes a kind of dependence between these two sets of considerations: how something about a person comes to be or maintains itself, and what could be seen to be reasonable, intelligible, or worthwhile in it. And finally, there is a further difference in how these two types of question 'why?' are treated, and that is that for the broadly normative type of 'why?' question as applied to someone's attitudes, we typically do ask that question of the person herself. Unlike the question 'why?' concerning a person's pain or sensation of vertigo, we take the person herself to be uniquely relevant as the person to ask regarding what is reasonable or worthwhile in some attitude she holds. 'Uniquely relevant' or 'indispensible' here does not mean infallible or incorrigible, but rather that if she cannot tell us why, then we may begin to doubt whether there is a good answer to that 'why' question. This is not an assumption we make about the question 'why?' concerning sensations or other bodily conditions, when the person has no answer of her own.[5]

When philosophers claim that the notions of reason and justification are internal to the notion of belief, this is not meant to deny that children and animals can have beliefs, even though they do not themselves have these notions of reason and justification. To say this much is so far just to say that certain norms of rational assessment apply to beliefs. This is not yet different from the idea that, e.g., it is internal to something's being a heart that it pumps blood and can be assessed as healthy or malfunctioning, even when the creature with the heart has no conception of such things. Controversy begins with the thought that, in mature humans, it belongs to the notion of belief that reason and justification not only *apply* to it, as a form of normative assessment, but that the believer play some role with respect to justification, a role that a creature does not typically play in the good functioning of its heart. The believer can be asked for her reasons for a certain belief, and the believer typically recognizes the applicability of that question, even when unable to give any convincing reasons on that occasion. The rational relations among beliefs (entailment, consistency, etc.) are recognized to be relevant to one's entitlement to maintain one's belief. In these ways we hold the believer responsible for her beliefs in ways that we do not typically hold the person responsible for the condition of her heart.

We do not ask very young children and nonhuman animals for their reasons for believing something, and yet it does seem to many philosophers to be central to the very notion of belief in mature humans that believing something opens one to the norms of justification, and the responsibility for conforming to those norms. We ask

[5] Compare Daniel Stoljar's discussion (this volume) of the distinction between "explanation-seeking" and "evidence-seeking" versions of the question "How do you know?"

the believer herself for her reasons, and the believer recognizes that she is indeed the person to ask, that the request for reasons is properly addressed to her. We do not address similar questions to the person about the condition of her heart. We do not have the same expectation that she will be the person in a special position to know about its condition and address the question of its good functioning. If this much is true, it raises several questions. One, what are we presupposing about the believer's knowledge of her belief and its justification when we address such questions to her specifically? Two, what are we presupposing about her agency with respect to her belief when we hold her responsible in these ways, different from any responsibility she may have for the condition of her heart? And three, how does this system of reason-asking and reason-giving among mature believers relate to the capacities for belief among young children and nonhuman animals? That is, if we grant that such creatures do have beliefs, and yet do not have the concepts and capacities that would enable them to engage in the system of reason-asking and reason-giving proper to mature believers, then what is left to the claim that it is somehow internal to belief among mature believers that the system of reason-asking and reason-giving belongs to it, and that the role of justification is recognized as such by the believers themselves?

In favor of retaining some version of the idea that the asking and giving of reasons belongs to the nature of belief and other attitudes themselves, I will just say the following. While some philosophers have gone so far as to doubt that the concept of belief can apply to animals and to children before they are language users, no one would want to deny that they are both capable of action in a perfectly ordinary sense. But at the same time, it seems we also do not want to say that having reasons for what one does is something only added on at a later stage and does not belong to the idea of action itself. Actions are purposive and goal-oriented. Both the child and the adult may have reasons for reaching across the table toward a glass of milk. The ability of the person to tell us what she is doing and why she is doing it is something that develops later, as part of the growth of various capacities and the initiation into various forms of responsibility. But the fact that rational assessment applies to the action, and that later these forms of assessment can be posed as questions that we direct to the person herself, asking her just what she is up to and why, is all a development of the same idea of action. These first-person capacities and responsibilities are no more extraneous to the idea of action than is the related capacity of a speaker to tell us what she means by something she said. At an early stage of the ability to talk, the child will not be expected to tell us what she means by her words. That is also something that develops later. But it is surely internal to the development of the child's very capacity as a speaker that eventually she is understood, by herself and others, to be in a special position to tell us what she

means, what she is talking about, and will not count as a speaker absent any such capacity. Here again, we need not suppose that a speaker has unbounded authority over the question of what she is saying, or that allowance cannot be made for one form of semantic externalism or another, but the fact that we can and do ask the speaker to clarify what she is saying is surely part of the very notion of saying something. In all these cases then, it will only in very special circumstances makes sense to say "Why ask *me*?" in response to a question about what one is doing, or what one is saying, or what one believes and why.[6]

If the second, normative version of the 'why?' question is also, like the first one, meant to shed light on how something comes to be, then it may be asked why it is that we take the person herself to be particularly relevant as the person to ask here. Why isn't another reasonably well-informed person an equally good or better source of information on this topic? And this question suggests the possibility that the reason we ask the person herself is closely related to the fact that the normative 'why?' question contains within it both a question about a form of normative assessment and a question about how something comes to be or maintains itself (the belief, the action). The condition of a person's heart can be normatively assessed by a doctor, but that condition is what it is quite independently of the person's cognitive relation to it, and it is not the person herself who is the locus of that assessment. By contrast, a person who believes it will rain, or who hopes it will not, is not just in a condition that can be normatively assessed, but is herself engaged in her own forms of normative assessment, and the believing or the hoping themselves are what they are in virtue of the person's overall sense of what would support or undermine them as attitudes. When the belief or the action is judged or found wanting, this is an estimation of the person, and not simply of some condition she is in. If attitudes such as these can themselves be seen as forms of normative engagement on the part of the person, then the question 'why?' is applied to them in a particular form. Part of what is meant in calling the believing itself a form of normative assessment is simply that to believe *p* is to take *p* to be reasonable, believable,

[6] This part of the story clearly presumes a creature not only capable of attitudes such as belief, but with the *concepts* of these attitudes. At an earlier stage of development, the child may be what Fred Dretske (this volume) calls an "unwitting authority" with respect to her beliefs. That is, she is aware of what she thinks, and she is the source for any claims about what she thinks, but, lacking the concept of belief, she cannot be said to know that she believes what she believes. There is undoubtedly a complex story to tell about the difference made to the character of the child's belief, and the possibilities for self-knowledge of her belief, when she comes to acquire the concept of belief itself. I take it that acquiring the concept will go hand in hand with such things as being asked "When you ran to the door, did you think Daddy was home?" "Why did you think so?" etc., and hence with the primitive forms of reason-asking and reason-giving.

and in more articulate contexts, defensible or justifiable. And although we do not choose our beliefs and do not perform them like actions, this relation to forms of normative commitment is a matter of common form between beliefs and actions: to believe *p* is to take *p* to be believable and open oneself to the question "why believe *that*?" and to do something is to take the action to be worth doing in some way, and thus to open oneself to the question "why are you doing *that*?" And in both cases the person takes the answer to the normative 'why?' question to be directly relevant to the existence or continuation of the belief or action in question. With this much in place, we can see it as making a certain sense to see the person herself as the one to ask when we want to know why she believes this or hopes for that, and why she will indeed recognize herself as the right person to ask, and might indeed insist that treating her relation to this question as something in principle dispensable, or at best one among indifferently many equally good sources of information, would be to fail to take her seriously in a fundamental way.

We have special reason to ask the person herself both what she is doing and why, or what she believes about something and why, because it is possible to see both actions and attitudes as themselves responses to questions of one form or another,[7] or as ways of resolving oneself. A belief is the answer to the question about what the evidence points to, or what best explains some happening; and an action expresses one's resolution with respect to the question what is to be done in a certain situation, or in response to a certain problem. As answers or rational responses, beliefs and actions invite the normative question 'why?' and assume the responsibility of the person to be able to speak to that question. By contrast, another internal condition of mine, like the condition of my heart or my sensation of pain, belongs to a different category and is not a possible answer to a question, or the possible conclusion of some line of reasoning (practical or theoretical). I may be responsible in one way or another for either the condition of my heart or my sensation of pain, but that is something purely external to the heart or the sensation itself, whereas on this view actions and attitudes are modes of resolving oneself, and hence involve forms of responsibility.

I take these considerations to amount to a reason to think that the ordinary person's ability to say what she is doing or what she thinks about something, and to know this without having to make the kind of observations of herself that she would if the question were about someone else's belief or action, is related to the fact that she bears a certain responsibility for her belief and action, fundamentally different from the responsibility a person may have for the sensations she finds herself with. In the case of action, we ask the person herself what she is doing because

[7] Pamela Hieronymi has stressed this formulation in recent work, although I suspect my understanding of the connection differs from hers.

we ordinarily take her to know, and we take her to know what she is doing because we take her to know why she is doing it, and we take her to know what she is doing and why because we expect this of her, it is her business to know. This brings a certain notion of agency into the picture of self-knowledge in that on this picture the non-observational character of self-knowledge with respect to actions and attitudes is tied to their being expressions of the rational, active side of one's nature. Hence, the relevant notion of 'activity' belongs to the *category* (attitudes, or actions themselves) as distinguished from another category (sensation or bodily condition), rather than to one's relation to a particular item in that category. Both a sensation of pain or a belief about my chances of winning the next hand of blackjack can be controlled or manipulated by me in various ways, and with various degrees of success. This notion of 'control' applies just as well to my relation to the perfectly inert objects in my immediate environment and is not relevant to the notion of agency being appealed to with respect to my doing and believing. Rather, if there is anything to this difference in category, then believing, intending, and hoping are themselves forms of activity, or expressions of the person's active nature. A person's beliefs are not chosen by her, nor are they typically "controlled" by her. Rather, what we call a person's beliefs are the precipitate of her ongoing rational activity. It is only derivatively that a person is 'active' with respect to a particular attitude itself. In the normal case, I find myself wanting to learn Russian, or suspecting that my relatives will not be visiting for the summer after all. Such attitudes do not emerge out of nowhere, of course, but rather become mine in the course of my ongoing thinking and acting, and are not aimed at as states to put myself in. I do not aim at acquiring some particular attitude, and its rationality is not expressed by my singling it out for control or manipulation. But for all that, my wanting, suspecting, and caring about something are expressions of my active nature, to which some form of the normative 'why?' question naturally applies, along with my taking myself to be the person who is answerable for why I do the things or believe the things that I do.[8]

3. Self-Knowledge and Settled Beliefs

Various recent writers have taken the 'transparency' of belief to be part of the explanation for the person's ordinary ability to say what she believes about something without having to base what she says on empirical observation of herself, but have found the appeal to rational agency to be misplaced. For some, this appeal has been

[8] For more on the distinction between exerting external control over one's attitudes and assuming rational responsibility for them, see Moran 2002.

thought to falter on an ambiguity between how I know what I *already* believe about something and how I know what I believe about it *now*, upon considering the question. Thus in a recent article, Nishi Shah and David Velleman say the following:

> The question "Do I believe that *p*?" can mean either "Do I already believe that *p* (i.e., antecedently to considering this question)?" or "Do I now believe that P (i.e., now that I am answering the question)?" . . .. . . Now, either of these questions can give way to the question *whether p*. If the question is *whether I already believe that p*, one can assay the relevant state of mind by posing the question *whether p* and seeing what one is spontaneously inclined to answer. In this procedure, the question *whether p* serves as a stimulus applied to oneself for the empirical purpose of eliciting a response. One comes to know what one already thinks by seeing what one says—that is, what one says in response to the question *whether p*. But the procedure requires one to refrain from any reasoning as to whether *p*, since that reasoning might alter the state of mind that one is trying to assay. Hence asking oneself *whether p* must be a brute stimulus in this case rather than an invitation to reasoning.[9]

This is said in the context of seeking to vindicate a notion of the transparency of belief, but it is not immediately clear just what sort of problem they see here, or

---

In "When We Are Ourselves: The Active and the Passive," Joseph Raz makes the case for seeing belief and related attitudes as part of our active nature in terms that are congruent with the account given here. I encountered his paper after completing this chapter, but the clarity of his account is worth quoting at length, particularly for the distinction between agency as responsiveness to reasons and the agency involved in explicitly arriving at a new belief. "We are active when our mental life displays sensitivity to reasons, and we are passive when such mental events occur in a way which is not sensitive to reasons; or at least this is part of what accounts for the distinction. In these terms beliefs are—pathological cases excepted—on the active side of our mental life. This does not mean that we form beliefs only as a result of deliberation. We may form them because—with our senses—we perceive how things are, or because through subconscious processes we come to have or to form them. All of this is consistent with the active character of believing, or having beliefs. . . . Even when we form perceptual beliefs, or when we come to have certain beliefs without being aware of the fact, the beliefs are responsive to reason. This responsiveness is manifested in two ways. First, in that unconscious processes of belief formation, just like explicit deliberation, depend on absence of awareness of reasons against the belief, and—normally—on reasons for it. When it seems to me that I see a cat I—without deliberation—believe that there is a cat there. But if I believe that I am in a magic show, then I do not form that belief. Second, when I deliberate and come to the view that the evidence is that a proposition that I believe is false the very process of coming to that conclusion is also a process of ceasing to believe it. By their responsiveness to reasons believing and beliefs belong to the active side of the active/passive divide" (1997, 218).

[9] Shah and Velleman 2005, 16. A related objection is made in the article by Alex Byrne (2005) discussed in the next section, as well as in Gertler (forthcoming).

how the kind of insulation from rational agency they have in mind could still deliver an answer about what I already believe about something. Does engaging my rational capacities corrupt the process of reporting on what I already believe about something? As a first approach, we could adapt an example from Sydney Shoemaker and suppose that I am asked who I believe was the President of the Confederacy during the American Civil War (Shoemaker 2003). It may happen that in responding to the question I start to say "Robert E. Lee" and then correct myself and say "Jefferson Davis." In this way, I failed to treat the question as a "brute stimulus," and in correcting myself I engaged my rational capacities, but none of this provides a reason for thinking that I have thereby produced a new belief rather than reported what I believed all along. Surely it may be the case in an example like this that I did in fact believe all along that it was Jefferson Davis, but blurted out the wrong answer and corrected myself, all the while being faithful to what I already believed. The engagement of my rational capacities in delivering the answer need not be seen as substituting a new belief for what I already believed.

On the other hand, there are difficulties in seeing how my response to such a question about what I believe could be a response to a "brute stimulus" as imagined here, and still be seen as reporting on a belief of any kind. To begin with, let us recall that with respect to any belief of mine, it counts as a belief insofar as I take it to be true (this, of course, is what makes some sort of appeal to 'transparency' seem attractive). If I relate to my "stored" belief as something I take to be true, it will be hard to see how I can see my relation to it, however spontaneous, as insulated from the engagement of my rational capacities for determining what is true or false. It cannot, for instance, be seen by me (or my auditors, if they are the ones applying the stimulus) as simply some name that is produced upon receiving the stimulus, for it has to represent what I take to be true, as an answer to the question asked. Hence, there is a considerable background involving my rational agency that has to be assumed for my response to the stimulus to count as my spontaneous answer to the question. Consider the fact that there are countless words and names that may be floating around in one's mind, and any one of them might be what comes out when the stimulus is applied. None of them will count as indicating one's belief about the matter unless, minimally, one can recognize the word as a name, and not something else, and recognize it as the name of a person, and recognized the name of this person to be relevant somehow to the stimulus such that it can serve as indicating what one's belief is about this matter. Simply hearing oneself coming out with something in response to a brute stimulus will provide no more reason for thinking this represents one's belief about something than if one were to sneeze in response to the stimulus. Rather, for my response to the stimulus to be seen as telling us what I already believe about the question, I have to *relate myself* in various ways to the

name I come out with, and not just hear myself say it. I must, at a minimum, understand the words I am saying, and understand them as responding to a question whose meaning I understand. And 'response' here must mean something like '*replying* to the question', for in a broader, more neutral sense of 'responding' we would need an additional reason to think that this 'response' bears any relation to my beliefs about anything (which is what the stimulus is supposed to deliver). For any mental content, word association, or exclamation that may be produced by the stimulus, it will count as relevant to the question of my belief only if I am relating to that content or word as representing my belief about the matter. That however, will mean engaging with my rational capacities in a way that involves my reflecting on the facts of the matter, which was supposed to be excluded from the process because it was thought to "contaminate" it.[10] But if that were so, then it is hard to see how a person could ever perform the ordinary task of telling us what they already believe about something.

What is meant by the "activity of reason" here need not be explicit, nor need it involve the production of a new belief. I may hear the question "Where was Balzac married?" and come out spontaneously with the name "Berditchev," and when asked why, perhaps I cannot say much more than "I must have read this somewhere" (see Dennett 1981, ch. 16). But to know that much about my response is already to know a great deal, is already for me to have classified my 'response' as the name of a place, in answer to a question about a person, and in connection with an event that left some written record I could have encountered. The confidence with which I spontaneously come out with the answer is surely dependent on such things as that I understand the statement proposed to me, that the proposition in question makes sense to me, and that the possibility it presents seems perfectly plausible to me, even if I can presently see no special reason to think it true. (Compare: the answer I come out with is "on the moon.") All of this and more is part of my apprehension of the rational environment of this proposition. There is no isolating my response to the question from all of this, and still have me responding to a question. It is only when we take for granted that this background is in place that we then confront the scenario of my replying to the question, with perhaps nothing more to go on than a feeling of familiarity, or the sense that I must have heard this before, but by then the work of reason has already prepared the place for my answer. We can, of course, isolate my response from reason if I do not understand the language in which the question is posed, or if the response I come out with makes no sense to me, something I just find myself saying, or if I come out with it as a sheer guess, but

[10] "As we pointed out, one cannot engage in reasoning aimed at answering the question whether *p* if one wants to find out what one already believes, because such reasoning would contaminate the result by possibly altering the state that one is trying to assay" (Shah and Velleman 2005, 17).

in none of those cases will we think that this response expresses my belief about anything. And with respect to other things I already believe, which are more integrated in the rest of my life and thought, things are more complex, and it is even harder to imagine what could be meant by isolating my response from the influence of reason. That is, if the question about what I already believe is something like: "Do you believe there are people living in Phoenix?" or "Do you believe that you can buy food with money?" then my spontaneous answer will be 'yes', not because I have reasoned my way to this conclusion, but because so much else of what I believe would have to be upended if this were not true, and I would have no idea how to begin such a revision of my beliefs. I do not have to think about it, but not because my answer is insulated from the influence of reason, but precisely because my answer is so fully integrated with the rest of my beliefs and rational capacities.[11]

What seems to be imagined in the scenario from Shah and Velleman is a situation of wanting to know what someone thinks about something, but prior to having that person consider the question itself. If I am playing poker with someone, I may want to know whether he thinks I am bluffing, but without actually raising that question with him, for that would risk alerting him to a possibility that I do not want him to consider if he is not doing so already. In this sort of case, I might do various indirect things to elicit information on this point, without raising the question itself. This might be thought of as a "brute stimulus" in the relevant sense, but it will importantly not involve my poker partner responding to a question as a "brute stimulus." Now, in my own case, it is much harder to imagine what would count as my "wanting to know what I think about some possibility, but without my considering the possibility itself." In the two-person case, I can raise the question of what this other person believes about *p*, without that person considering the question of the truth of *p* itself. In my own reflections about him, I can speculate about what may or may not be going on there in this other mind, knowing that my reflections are strictly mine, and not part of what constitutes the state of mind I am speculating about. But with only one person on the scene, this is not a real possibility. I cannot pose the question to myself of whether I believe that *p* without raising the question of the truth of *p*, for there is only one mind under consideration here, inquiring about itself. Naturally, this does not mean that I cannot, in certain circumstances, seek to "assay" what I really believe about something, in a way that

[11] See Wittgenstein 1956, § 478:

What kind of reason have I to assume that my finger will feel a resistance when it touches the table? What kind of reason to believe that it will hurt if this pencil pierces my hand?—When I ask this, a hundred reasons present themselves, each drowning the voice of the others. "But I have experienced it myself innumerable times, and as often heard of similar experiences; if it were not so, it would. . . . "; etc.

brackets the question of the truth of what I seem to believe. In various circumstances, a person can indeed take such an 'outsider's' perspective on her own belief, even though the result may be an inherently unstable one ("Well, I know that the plane is safe, but clearly I am also in the grip of a fear that it is not safe.") But even this possibility is quite different from the idea that I could raise the question of whether I really think the plane is safe without considering the question of whether the plane is safe. I have to understand what question I am asking myself (the way my poker partner does not have to understand my question about him); or if I am applying a 'stimulus' to myself I have to know how I understand the relation between the stimulus I am applying and the question I am seeking to answer. And I cannot understand either of these things without considering the *content* of the state of mind I am inquiring into, which orients me with respect to the question of its truth, reasonability, or comprehensibility. Inquiring into one's own mind about what one already believes about something cannot be insulated from one's rational agency in the way that it can be with respect to the mind of another person, but nor does this involvement mean that what I already believe is elusively out of my reach, because always threatened with replacement by something else once my reason is engaged. Unlike my relation to my partner in poker, in the first-person case, inquiry into the mind has to be intelligible to the very mind being inquired into, and hence involves the engagement of reason on the part of both the inquirer and the mind inquired into, for they are one.

4. Transparency and Rules for Belief

In a recent paper, Alex Byrne (2005) seeks in a different way to account for Transparency, while avoiding appeal to notions of activity or rational agency in the explanation of what makes appeal to Transparency possible. The paper presents a rethinking of several issues concerning self-knowledge, and its general aim is to vindicate what Shoemaker criticizes under the name of the Broad Perceptual Model of self-knowledge. This is to be distinguished from any appeal to "inner sense" (although Byrne has his sympathies here[12]) and comes down to the two claims that (1) our detection of our mental states is based on some sort of causal mechanism, and (2) that our mental states themselves obtain independently of our access to them (Byrne 2005, 86). Shoemaker's case against the "broad perceptual

[12] To be sure, much of the paper is devoted to showing how the case against "inner sense" turns out to be unconvincing, and hence counts as a defense of that view. But at the same time "inner sense" is described as an "extravagant" rather than "economical" account of self-knowledge, in that

model" depends on challenging (2) rather than (1), since Shoemaker (this volume) defends the idea of a "constitutive relation between believing something and believing that one believes it." The constitutive relation between one's mental life and one's first-person access to it is defended by Shoemaker as a consequence of being in a certain belief state, combined with ordinary rationality.[13] In this sense, then, although we do not infer or reason our way to knowledge of our beliefs, there is a constitutive connection between being a subject of belief in the first place and having one's beliefs available to one, without the need for the kind of evidence one would need for knowledge of the beliefs of others.

Byrne's alternative claim is that the phenomena of self-knowledge can be understood as the result of our following a rule for belief formation, which he calls BEL. The phenomena to be accounted for include both privileged access, described as the idea that our "beliefs about our mental states acquired through the usual route are more likely to amount to knowledge than beliefs about others' mental states" (2005, 80), and what he calls "peculiar access"; that is, the idea that one has a special method for learning of one's own beliefs that cannot be applied to the beliefs of others (ibid., 81). The BEL rule is presented as a reconstruction of the Transparency condition, and an explanation of why it is legitimate for a believer to answer a question about her belief about something, by reflection not on herself, but on the topic of the belief in question. Hence, he presents the BEL rule as follows:

BEL: If *p*, believe that you believe that *p*.

This rule for belief is introduced by comparison with another rule, called DOORBELL, which states:

DOORBELL: If the doorbell rings, believe that there is someone at the door.

DOORBELL is what Byrne calls a "good rule" insofar as it "tends to produce knowledge about one's visitors" (ibid., 94), and this goodness will depend, naturally, on various empirical conditions obtaining in the environment of the person following this rule. It can lead one astray, and hence is not failsafe, but may be reliable enough in practice to count as a rule the following of which results in knowledge. BEL, however, has epistemic virtues superior to DOORBELL in that it is "self-verifying": if it is followed,

---

it requires appeal to an additional capacity or mechanism, beyond what is already needed for our general capacity for rationality (Byrne 2005, 92). The account Byrne later goes on to defend, in terms of the BEL rule, is recommended for being economical rather than extravagant (ibid., 99), hence I take it that Byrne should be understood as defending a version of the "broad perceptual model," but not "inner sense." "The account is not a version of the inner-sense theory" (ibid., 9).

[13] "[B]elieving that one believes that *p* can be just believing that *p* plus having a certain level of rationality, intelligence, and so on" (Shoemaker 1994, 244, quoted by Byrne on 89).

the resulting second-order belief will be true (p. 96). This has the nice feature of also capturing "peculiar access," since BEL will only be self-verifying when applied to one's own beliefs. The variant of BEL "If *p*, believe that Fred believes that *p*" is not a good rule. Finally, BEL is an especially "safe" rule in that the resulting second-order belief will be true, even if one tries but does not succeed in following it. As the rule is formulated, actually applying it requires the obtaining of some fact '*p*', hence in the absence of that fact the BEL rule is not in fact being followed. However, BEL has the special virtue that even trying to follow it will result in a true second-order belief, since if I mistakenly take '*p*' to obtain, and then seek to follow BEL on the basis of this mistake, I still end up with the true second-order belief that I *believe* that *p*.

BEL is described as a rule that we follow, and the following of this rule is described as form of *reasoning* (ibid., 94). And indeed, the following of this rule is described by Byrne as no more problematic than the following of less exotic epistemic rules, such as DOORBELL.[14] At the same time, Byrne does acknowledge BEL as an unusual epistemic rule since, unlike DOORBELL, it recommends a transition from the apprehension of a fact to a belief concerning a logically independent fact, without the first fact being evidence for what is believed in the resulting belief. One way to display the difference between the two rules is by noting that, despite the imperative form in which it is presented, DOORBELL can also be formulated as simply a relation between two contents, rather than as instruction for belief formation:

DOORBELL*: If the doorbell rings, there is probably someone at the door.

And indeed, upon hearing the recommendation of the original DOORBELL rule, it would be natural to assume that the only reason for following it was that there was just such a relation between the two contents, such that the fact of the ringing supported or was evidence for the fact of someone's being at the door. Absent such a connection between the contents in the two parts of the rule, it would be difficult to embark upon complying with it, unless one could somehow install beliefs in oneself, exerting a kind of purely pragmatic agency with respect to one's beliefs. As mentioned, Byrne is well aware of this difference between BEL and DOORBELL, since in a sense this difference simply comes to the original puzzle of Transparency: How can it be legitimate to answer a question about a particular person's beliefs, by appeal not to facts about that person, but to facts relating to the content of the belief? The parallel with the revised DOORBELL rule, in terms of a relation between two contents, would be:

[14] "Given that we follow rules like DOORBELL, it should not be in dispute that we *can* follow BEL" (ibid., 96).

BEL*: If *p*, then I probably believe that *p*.

As Byrne points out, however, "this is a *bad* rule: that *p* is the case does not even make it *likely* that one believes that it is the case" (ibid., 95). How then are we to understand following the original BEL rule as a form of reasoning when it does not present a rational relation of support between the two contents? If BEL is a rule we can be said to follow, and the result of following it is the formation of a particular belief, then there must be some answer to the question of the person's entitlement to make the transition in thought that is being recommended by the rule.[15] Whatever "following the rule" comes to, it cannot be the exercise of a kind of agency unrelated to the truth-centered demands of belief (as it would be for instance with the recommendation, "Believe that *p*, because then you will stop worrying so much that not-*p*"), for the account is meant to be an account of self-knowledge. The fact that BEL is "safe" in the sense of being unlikely to deliver false beliefs would be a recommendation of it as a description of a good mechanism of belief formation, but the person seeking to follow the rule, the way she follows DOORBELL, will still be in need of some reason relating the two contents. Otherwise the rule is reduced to saying, as it were, "Act upon yourself, do whatever it takes to produce the one belief on the basis of the other."

On externalist epistemic assumptions, one may be entitled to some belief without having reasoned one's way to it, and without now being able to provide justification for it, so long as there is in fact a reliable connection between one's belief and the fact in question. A reliable mechanism for the production of true beliefs need not represent itself in terms of a rational connection between belief contents, but may instead operate without any understanding of it on the part of the person herself. BEL, on the other hand, represents itself as a rule for belief, expressed to the believer as an imperative or a recommendation. As such, BEL is not just the description of a good disposition for someone to have, but rather requires the believer to do something in response to it, and in accordance with it. This much is needed by the parallel with DOORBELL, and the understanding of BEL as itself a form of reasoning. Following a rule for belief, however, is not simply undertaking to produce a belief in oneself by whatever means necessary, but requires from the rule follower some understanding of, and an endorsement of, the rational connection between the contents mentioned in the rule. DOORBELL does provide this, given the relation between the contents as represented in DOORBELL*, but BEL itself does not.

[15] I speak of "transitions" of thought here so as not to pre-judge the question whether we should understand BEL as an inference, as a rational connection of another sort, or as a kind of "blind" rule that nonetheless produces "good" results ("safe" beliefs). Shoemaker (this volume) questions whether following the BEL rule can be understood as a form of reasoning.

Why should there be such a difference in apparent "goodness" between the original BEL rule, an imperative or recommendation addressed to some "you," and the revised rule BEL* relating two contents, especially when the corresponding two versions of DOORBELL do not display this difference in goodness, and indeed seem to support each other? And can the imperative form retain any force when divorced from the relations of support among belief contents described in rules like DOORBELL*? Looking at BEL as an account of Transparency, and as a proposed answer to the problem of "two topics" (e.g., the weather, my belief), we might look at the issue in terms of three possible candidates for belief and their requirements.

1. Considering *p* as a candidate for belief, I require evidence for *p* or truth-centered reasons of some sort if I am to believe *p*.
2. Considering the candidate "Jones believes p," I likewise require evidence or other truth-centered reasons concerning *Jones*, since he is the subject of the content in question.
3. However, considering "I believe that *p*" as a candidate for my own belief; that is, the question of attributing the belief *p* to myself, I do not appeal to evidence for the content "I believe that P," where that is taken to refer to the beliefs of a particular person, as in (2). Rather, I appeal to the sorts of reasons mentioned in (1), reasons in favor of *that* content, the one that is embedded in (3), and which does not mention any person.

Statement (3) is a reconstruction of Transparency as I have been understanding it. As such it raises the question of what could make legitimate the appeal to reasons relating to the embedded content (P), rather than to reasons relating to the ostensible content of the attribution itself ('I believe that P'). On this way of looking at it, the beliefs governed by Transparency are not altogether independent of the appeal to evidence or other truth-centered reasons, but the question is what makes legitimate the exclusive appeal to reasons relating to the embedded content (1) rather than reasons concerning the apparent content of a psychological attribution (2). What breaks the apparent parallel between (2) and (3), such that (3) is answered as though it were a version of (1)?

As with the original Transparency condition, the appeal to the BEL rule also relates the question of the content *p* (1) to the question of the self-attribution (3), and hence incurs a similar burden to explain how it can be legitimate to make the transition from '*p*' to 'I believe that *p*'. The explanation adverts to the self-verifying character of following BEL, but that will only be a good explanation if we have a better understanding of what "following" it comes to, and the conditions under which doing so is possible. In particular it bears explaining how the imperative form of BEL could make sense to someone seeking to follow it when it is understood to

lack the support of the corresponding version of the rule (BEL*), which describes relations among the contents of beliefs ("If *p*, then X believes that *p*"). If we distance ourselves from the imperative form of the rule for a moment, we can describe the believer who is guided by the BEL rule as someone who makes a transition from the apprehension of some fact *p* ("It's raining out") to a belief about herself. This "belief about herself" is the product of following the BEL rule, which the imperative form presents as addressed to "you" ("Believe that *you* believe that *p*"). In taking it out of the imperative form and representing the transition in thought that the rule recommends, we need some substitute for "you," and it will be crucial that the substitution retains the status of BEL as a "good rule." If, for instance, the description of the transition takes us from "*p*" to "Jones believes that *p*," this will be a bad rule, even if the person making this transition is indeed Jones. For he may not recognize himself as "Jones," in which case the BEL rule would be taking him from the apprehension of some fact "*p*" to a belief about the beliefs of some person named Jones. The problem of "two topics" would re-emerge for the understanding of Transparency, the problem of understanding what could be legitimate in the transition from a thought about, e.g., the weather, to a thought about some person's beliefs. Hence, the transition described in the BEL rule is subject to certain conditions, familiar from the philosophical discussion of the first person. The transition described in BEL will only be legitimate when the person following it is identical to the person whose beliefs are mentioned in the resulting second-order belief. And this identity must be recognized by the person making this transition, and hence the name we substitute for "you" when we describe this transition outside the imperative form of the rule must be one under which the person in question could not fail to recognize herself. If there were a possibility of misidentification or failure to identify oneself with the person whose beliefs are the topic of the resultant second-order belief, the application of the BEL rule would be as illegitimate as the transition from some proposition *p* to the beliefs of some arbitrary person.

The 'I', then, is what lies behind the 'you' who is the implicit addressee of BEL in its imperative form. And in particular the "subject use" of 'I' is crucial to understanding how it is that, while the attribution of belief to a person under some name, description, or demonstrative must appeal to evidence concerning that person (as in (2)), the first-person discourse of belief appeals to reasons concerning the content of the belief rather than evidence or identifying information about the believer.[16] When the BEL rule instructs someone to "believe that you believe that *p*," this "you"

[16] For the initiating discussions of the idea of this use of 'I' proceeding independently of "identifying information" about the person, see Wittgenstein 1958, Shoemaker 1968, and Evans 1982, ch. 7.

must not be in need of identifying information anymore that the subject use of "I" is, if the following of the rule is to be safe or self-verifying. If identifying information were required to pick out the right person named in the instruction "believe that *you* believe that *p*," or if evidence about that person were needed to determine what that person's beliefs were, then forming the belief about the beliefs of this person "you" would be in need of the same kind of support, and subject to the same semantic and epistemic risks, as would the corresponding belief about the beliefs of Jones. Under these conditions BEL would be "bad" in the same way noted by Byrne in connection with the "neutral schematic" version of BEL: the obtaining of some fact *p* would be treated as a reason for concluding that some person believes it (ibid., 95).

So a fuller picture of the conditions under which BEL would be a good rule are that (a) it is only knowledge of one's own beliefs, and not those of another person, that following BEL will make possible, and (b) that in this knowledge I *recognize* myself as the person whose beliefs are the topic of this knowledge, and (c) that my recognition of myself as the person whose belief is known in this way not be based on *identification information* of any sort. One situation in which a person may indeed need to avail herself of identification information regarding some attitude is when she takes a belief or a fear of hers to be evident in her behavior but otherwise inaccessible to her because it is not responsive to her sense of the reasons that would support the belief or fear. If I have reason to believe that what sustains my actual attitude regarding someone's competence or trustworthiness is some fear or prejudice irrelevant to the question of his actual competence or trustworthiness, then I will not take myself to be entitled to make the transition described in the Transparency condition or the BEL rule. If I think that my actual belief about him (or people "like him") is controlled by prejudices whose operations are beyond my awareness and which persist independently of my grasp of the reasons for or against *p*, then I will not take myself to be entitled to answer the question "Do you believe that *p*?" by reflection on the question '*p*?' Likewise, I will not be entitled to follow the BEL rule, which tells me that if "He is just as competent as the others," I may help myself to the belief that I believe he is just as competent. The difference between the topics of the two questions will re-emerge. I will recognize that I may need to give different answers to the two questions. I may know about this tendency of mine, and know that the beliefs that are its product are resistant to change, and may not be reflected in my overt, sincere declarations when considering the content itself.[17]

[17] See Moran 2003, 407–408, for more on different ways of imagining such failures of transparency.

Thus, the appeal to either Transparency or the BEL rule has various conditions, which matter to the understanding of the kind of rationality that is assumed in making the transition from '*p*' to "I believe that *p*." For a person to be entitled to make this transition, he must recognize the identity between himself and the person whose belief is the topic of "I believe that *p*," and he must take himself to be able to speak for the beliefs of that person, in the sense that he presents himself as the very person whose sense of the reasons in favor of '*p*' are expressed in the statement "I believe that *p*." Transparency fails if the statement I arrive at, "I believe that *p*," is delivered in a purely *attributive* mode, for that would require evidence about me, which Transparency does not provide.[18] What Transparency does purport to provide is a connection between my reflection on a particular content '*p*' and the reasons in favor of it, and the answer to the question of what my actual belief is. That connection presumes that my belief about some matter '*p*' is determined by my sense of the reasons in favor of '*p*', and not by forces independent of those reasons. This is not a matter of constructing a new belief, but of seeing my past settled beliefs as available to me through Transparency insofar as they make sense to me in the light of reflection on their contents.

This requirement does not mean that the declaration "I believe that *p*" cannot be the expression of what one already believes,[19] but that the statement of belief must be delivered in the mode of endorsing the content of the belief, and not as the attribution of belief (to oneself, to another person) in a way that leaves open the question of endorsement. The idea of endorsement points to a form of rational agency that both Byrne, and Shah and Velleman, find misplaced in the account of Transparency, but in distancing themselves from it, they each in their different ways picture the self as exercising the "wrong kind of agency" with regard to its own beliefs, either by seeking to elicit one's own beliefs through a kind of stimulus insulated from the exercise of reason, or by proposing the application of a rule of belief formation that could only be applied "externally," since we lack a rational connection between the two belief contents. The centrality of the first person to the understanding of transparency points to the fact that conforming to it should not be pictured as producing a new state in oneself so much as committing oneself on a certain matter.[20] The transition described in Transparency is not an inference from evidence about a particular person, but rather something more like a general

[18] On "attributive" see Moran 2001, ch. 3.3.

[19] Byrne (2005, 85). Gertler (2011) makes a similar criticism, but unlike Byrne, she makes it in the course of denying the appeal to transparency, rather than as a reconstruction of it.

[20] Boyle (forthcoming) argues for seeing the notion of agency relevant to self-knowledge as a form of "self-activity" and against what he calls the "process conception" of deliberation.

presupposition of rational thought, to the effect that, from the first-person point of view, I must take what I believe about something to be the expression of my sense of the reasons relating to the content of that belief. As we have seen, it is a presupposition that may lapse in cases of compromised rationality, but at the same time, it is hard to see what could be more basic to rationality than the idea that I take the question of what I believe about X to be determinable by my reflection on X itself. This is why Transparency is more like a *requirement* than like a permission or an optional means for determining one's attitudes, for if I could not learn what I (really) think about X by reflection on X itself, that would mean that my belief about X was out of reach of my reflection on the facts concerning X itself. If we assume a subject with mastery of the first-person pronoun, and its independence of identifying information, it will only be in conditions of compromised rationality that a person could believe that it is raining, possess the concept of belief, and yet be unable to know her belief through reflection on the weather.[21]

5. The Activity and Passivity of Belief

Part of the sense of 'rational agency' in this domain lies in the fact that, with respect to something I already believe, if I am considering it as something I am in fact committed to, then at a minimum I should be able to see my having this belief as a potential reason for someone else to believe the same thing. My confidence in some belief of mine may vary, but this will vary in accordance with whether what I take myself to discover is a belief rather than a hunch or a mere guess. Hence, I may not be prepared to assert it, but if it is a belief at all I must take the fact that I think so to be something that could be part of someone else's reason for believing the same thing. Hence, while I may not be 'active' with respect to it in the sense that it is a conclusion I have arrived at just now, seeing it as a belief of mine means such things as knowing what would counter it, knowing what sorts of reasons are relevant to it, my being prepared to increase or decrease my credence in it depending on other things I learn.

[21] Compare Byrne, in summarizing his account as a version of the Broad Perceptual Model: "Thus there is an appropriate causal mechanism. The state detected is independent of its detection. The subject might not have followed BEL, in which case the first-order belief would have been present without the second-order belief. What's more (we may fairly supposed) someone might believe that it's raining, possess the concept of belief, and yet not even have the capacity to follow BEL" (2005, 98).

More importantly, even for perceptual beliefs that are in some sense *forced* on me, these still differ from mere *feelings* (sensations) in that a perceptual belief is still something I am prepared to assert and defend.[22] Normally, I take my perceptual belief to put me in a position to claim something, something of potential epistemic import for others, and not just myself. Here we already see something of my agency involved in a way that we do not in considering my headache. In some sense I may be said to be passive with respect to both of these, but in the case of my perceptual belief the sense of my having no choice in the matter is precisely a matter of my finding the reasons in favor of the belief to be unassailable. This is not a matter of something merely befalling me, like a headache. Rather, my 'passivity' here stems from the fact that, for me, to change my belief would require a reason, and a reason of a special sort (viz., a reason connected with truth), and hence I do not have what Pamela Hieronymi calls 'discretion' with respect to this change, as I do with respect to the question of changing my shoes (Hieronymi 2006). So, in that sense, my 'agency' is quite restricted here. But that restriction is just the reflection of the dominance of another kind of agency, that is, my responsiveness to epistemic reasons for my beliefs, taking them to be the basis for what I can assert. And in the case as we are imagining it here, within this restricted range of admissible reasons there is dominance with respect to a single conclusion: I have every reason to trust the visual evidence here and no reason to doubt it, hence I cannot help but believe there is, e.g., a tomato in front of me. This is not like being assailed by a stabbing pain (something not subject to justification in the first place), nor is it like being carried away by a compulsion to believe that I will win the next round of black jack because the stakes are so high. Rather, it is an instance of what is called being compelled by the weight of reasons in favor of some conclusion. I have no choice because change of belief requires reasons, because beliefs are things I am prepared to defend, and in this case the reasons are overwhelming in a particular direction. This 'passivity', then, is an expression of my rational agency and is utterly unlike my passivity or lack of choice with respect to how fast my hair grows.

In this way, then, the 'passivity of belief' is the reverse side of a person's rational agency as a believer, for it is because one's beliefs are the expression of one's rational relation to the world that they cannot be simply "chosen." If what I believe were not answerable to the ways the world is, then I could indeed treat my beliefs as states which I could seek to produce in myself for reasons unrelated to their truth. I could "control" them in a way that enjoys "discretion" over the kinds of considerations that will count for me as reasons in favor of bringing it about. Likewise, when I am

[22] Here again, see Sosa (this volume), as well as the quotation from Raz in note 8.

relating to the beliefs of another person, I may also see them as states to be produced or removed for purely practical reasons of my own. In this sense I can be 'active' with respect to them the way I am when I seek to manipulate the objects in my environment. The fact that I am not active in this sense with respect to my own beliefs does not, however, mean that I am passive with respect to them the way I am passive in relation to a sudden pain or something falling on top of me. The very limitation in my manipulator's or producer's relation to my own attitudes is an expression of the fact that they are my cognitive and affective relation to the world, and hence demand from me reasons connected with the kinds of attitudes they are. A belief or a hope represents the world in a certain way, is something to which justification is internal, and which stands in logical relations to other attitudes. While there may be reasons for having a sensation, the sensation is not itself something reasonable or unreasonable. By contrast, the reasons that are primarily relevant to believing are not reasons for having the belief, but rather reasons in favor of the content of the thought, that is, reasons relating to its truth. The fact that in the case of attitudes there is a contrast between reasons in favor of having the attitude and reasons in favor of the content of the attitude itself is something that distinguishes them as a category from the category of sensations, for there is no corresponding contrast in the case of sensations.

When it is said that we cannot adopt a belief "at will," part of what this means is that one cannot adopt a belief on a whim, in response to a simple request from another person, or otherwise for reasons unrelated to its truth. Discretion of this sort represents a kind of freedom or liberty we are familiar with in other contexts, a form of freedom that appears to characterize our relation to our ordinary actions (e.g., changing one's shoes), but does not appear to apply to our relation to our beliefs and other attitudes. Hence, it may be natural to conclude from this that the notions of agency and activity have no bearing on the nature of belief and other attitudes, that they are "forced upon us," in much the same way as are the facts of gravity or the inert objects in one's environment. Or it may suggest that if a form of agency did characterize a person's relation to her beliefs, that could only be as a form of managing or manipulating a set of items within one's purview. This is, of course, not only unrealistic as a picture of psychological life, but neglects the fact that the great majority of one's beliefs and other attitudes manage to maintain themselves in a rough rational systematicity without any need for monitoring or intervention from the person as such. And hence it may seem natural to conclude from this that self-knowledge of attitudes like belief has no particular bearing on their rationality or on epistemic responsibility generally.

I have tried to suggest room for a different picture of the relation between self-knowledge and agency, which shifts the picture of self-knowledge from that of monitoring an internal condition to the ordinary ability to say what I am doing and

why, or what I think about something and why. This also involves a difference in the kind of agency invoked here, a shift from agency as production, to agency as responsiveness to reason. In one sense raising my arm voluntarily is an expression of my active nature, and in a quite different sense what I count as a reason is an expression of my active nature, as well as how I take myself to be answerable for what I believe. With respect to the beliefs and attitudes of another person, I may exercise a kind of productive agency, aiming at specific results. Here my reasons for favoring the production of the belief as state may be fully separate from reasons in favor of what is believed, and may be as diverse as my various reasons for wanting something to happen. In the first-person perspective on belief, however, my primary relation is not to the fact of having some belief but rather the commitment to its truth and what that requires of me. Detaching my relation to a state of belief (mine or another's) from the commitment to its truth is precisely what would allow for discretionary reasons in relation to its production. The first-person point of view presumes the absence of such separation, presumes the identity of the considerations in favor of the thing believed with the fact of one's believing it. The absence of such separation characterizes the kind of rational agency involved, specifically that it is not a matter of acting upon oneself or taking oneself as an object, but rather of being resolved about some matter. It is this same absence of separation that characterizes the kind of self-knowledge in question, namely that it makes possible a form of 'transparency', the fact that a person can normally tell us what she thinks about some possibility by reflecting on that possibility itself.[23]

REFERENCES

Bilgrami, Akeel. 2006. *Self-Knowledge and Resentment*. Cambridge, Mass.: Harvard University Press.

Boyle, Matthew. 2009. "Two Kinds of Self-Knowledge." *Philosophy and Phenomenological Research* 78(1): 133–164.

———. Forthcoming. "On Making Up Your Mind and the Activity of Reason."

[23] In writing this chapter I have benefitted profoundly from conversations with Matthew Boyle, and from his seminar on "Self-Consciousness and Self-Knowledge" in Fall 2009. I had very helpful comments from and conversation with Endre Begby, as well as other participants at the conference "Self-Knowledge and Rational Agency," sponsored by the Center for the Study of Mind in Nature, in Oslo, June 2010. I am grateful as well to Alex Byrne, Jonathan Vogel, Pamela Hieronymi, and Doug Lavin for conversations about these issues, and to the editors of this volume for very helpful comments on the final version, as well as the comments of two anonymous reviewers.

Burge, Tyler. 1998. "Reason and the First-Person." In *Knowing Our Own Minds*, edited by C. Wright, B. C. Smith, and C. Macdonald. Oxford: Oxford University Press.

Byrne, Alex. 2005. "Introspection." *Philosophical Topics* 33(1): 79–104.

Dennett, Daniel. 1981. *Brainstorms*. Cambridge, Mass.: MIT Press.

Edgley, Roy. 1969. *Reason in Theory and Practice*. London: Hutchinson.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Gertler, Brie. 2011. "Self-Knowledge and the Transparency of Belief." In *Self-Knowledge*, edited by A. Hatzimoysis. Oxford: Oxford University Press.

Hieronymi, Pamela. 2006. "Controlling Attitudes." *Pacific Philosophical Quarterly* 87(1): 45–74.

Moran, Richard. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press.

———. 2002. "Frankfurt on Identification: Ambiguities of Activity in Mental Life." In *Contours of Agency: Essays for Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, Mass.: MIT Press.

———. 2003. "Responses to O'Brien and Shoemaker." *European Journal of Philosophy* 11(3): 402–419.

———. 2007. Essay Review of *The Reasons of Love* by Harry Frankfurt, *Philosophy and Phenomenological Research* 74(2): 463–475.

O'Brien, Lucy. 2007. *Self-Knowing Agents*. Oxford: Oxford University Press.

Peacocke, Christopher. 1998. "Conscious Attitudes and Self-Knowledge." In *Knowing Our Own Minds*, edited by C. Wright, B. C. Smith, and C. Macdonald. Oxford: Oxford University Press.

Raz, Joseph. 1997. "When We Are Ourselves: The Active and the Passive." *Supplement to Proceedings of the Aristotelian Society* 71: 211–227.

Shah, Nishi, and David Velleman. 2005. "Doxastic Deliberation." *Philosophical Review* 114(4): 497–534.

Shoemaker, Sydney. 1968. "Self-Reference and Self-Awareness." *Journal of Philosophy* 65(19): 555–567.

———. 1988. "On Knowing One's Own Mind." In *Philosophical Perspectives: Epistemology*, edited by J. Tomberlin. Ascadero, Calif.: Ridgeview Publishing.

———. 1994. "Self-Knowledge and 'Inner Sense'." *Philosophy and Phenomenological Research* 54: 249–314. Page reference to the reprinting in Shoemaker (1996).

———. 1996. "Moore's Paradox and Self-Knowledge." In *The First-Person Perspective and Other Essays*, by Sydney Shoemaker. Cambridge: Cambridge University Press.

———. 2003. "Moran on Self-Knowledge." *European Journal of Philosophy* 3(3): 391–401.

Wittgenstein, Ludwig. 1956. *Philosophical Investigations*. Oxford: Blackwell.

———. 1958. *The Blue and the Brown Books*. Oxford: Blackwell.

Wright, Crispin, Barry C. Smith, and Cynthia Macdonald, eds. 1998. *Knowing Our Own Minds*. Oxford: Oxford University Press.