SELF-IDENTIFICATION AND SELF-KNOWLEDGE

by

HONGWOO KWON

B.A., Seoul National University, 1998

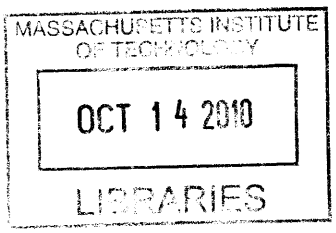
Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010



ARCHIVES

Signature of Author .....
Department of Linguistics and Philosophy
September 9, 2010

Certified by .....
Robert C. Stalnaker
Laurence Rockefeller Professor of Philosophy
Thesis Supervisor

Accepted by .....
Alex Byrne
Professor of Philosophy
Chair of the Committee on Graduate Students

SELF-IDENTIFICATION AND SELF-KNOWLEDGE

by

HONGWOO KWON

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

September 2009

Abstract

The traditional view has it that self-locating beliefs are distinctive in that they have distinctive contents. Against this, I claim that the distinctive element of self-locating beliefs should be placed outside contents. If someone believes that he himself is hungry, he not only has a propositional belief of a certain particular person that he is hungry, but also identifies himself as that particular person. The latter is not a matter of propositional belief, but a matter of taking a first personal perspective on that person's actions, beliefs and experiences. A subject takes his actions and beliefs to be "up to" himself, and regards his experiences as giving information about where he is located in the world. All these phenomena are shown to be related to the peculiar ways in which we come to know certain facts about ourselves. So self-identification is conceptually connected to self-knowledge. The three chapters discuss some parts or aspects of this reasoning.

Chapter 1, "Perry's Problem and Moore's Paradox," claims that Perry's problem of the essential indexical and Moore's paradox are essentially a single problem applied to two different aspects of our rational activities, actions and beliefs, respectively.

Chapter 2, "On What the Two Gods Might Not Know," defends what may be called an ability hypothesis about self-locating knowledge, drawing on David Lewis's ability hypothesis about phenomenal knowledge. What the gods might lack is best viewed as the abilities of self-knowledge.

Chapter 3, "What Is the First Person Perspective?" asks what it is to take a first person perspective and view oneself as the author of one's own actions. It is a matter of taking a deliberative stance toward one's own actions, which in turn can be best understood as the special ways in which we know them.

Thesis Supervisor: Robert C. Stalnaker
Title: Laurence Rockefeller Professor of Philosophy

Acknowledgements

Thanks to my thesis committee, Bob Stalnaker, Alex Byrne, and Rae Langton for their insights, patience, and encouragements. I am especially grateful to Bob, whose writing and advice have great influence on my philosophical thinking. I also thank the MIT philosophy program for its generous financial support, and its making an excellent environment for studying philosophy.

Thanks to my family and friends for their support and affection they have provided me over years. I am especially grateful to my parents and parents-in-law for their constant confidence in me. Lastly, my deepest thanks to my wife, Sunhyung Rhee, who has gone through all this time with me, for all her love and support.

Contents

CHAPTER 1

Perry's Problem and Moore's Paradox

1.1 Introduction

John Perry's problem of the essential indexical (or Perry's problem for short) and Moore's paradox are two familiar puzzles concerning thoughts about the first person.[1] I think the two problems are closely related.

Suppose that Perry says to me, "I have a class to teach in half an hour," and I come to believe what he says. We can easily anticipate what will follow; he soon will start moving toward the campus. It is because his believing this, with other background beliefs and desires, motivates him to move for the campus. But note that I also came to believe the same thing, which I will express by saying, "You have a class to teach in half an hour." Moreover, let us assume, I share other relevant beliefs and desires with Perry. But what I am motivated to do will be different; I may urge him to go. If Perry and I share all relevant beliefs and desires, why are we motivated to act differently?[2] The fact that Perry *identifies*

[1] Perry (1979; 2006) and Moore (1944/1993).

[2] Perry formulates the problem in a couple of different ways, and in this paragraph, I am following Perry (2006) (see, in particular, sec. 7). I should make it clear at the outset that Perry's problem as he originally formulates it explicitly and almost exclusively concerns the essentiality of indexicals in the contexts of *action explanation*. See Perry (2006), 213, for an

himself as the person who has a class to teach, which is apparently not captured by believed propositions, seems to make this difference. And Perry's use of the indexical "I" in expressing his belief seems to indicate this fact.[3]

Suppose that Moore says to me, "It is raining, but I don't believe it." It sounds absurd or nonsensical, so much so that it even makes me suspect that there is something pathologically wrong with him. But on a second thought, let us assume, I have believed the same thing, which I will express by saying, "It is raining, but you don't believe it." But if Moore and I say and believe the same, then why does Moore's saying or believing it look so bad, while my saying or believing the same is just fine? The fact that Moore *identifies himself as* the person who believes it, which is apparently not captured by believed propositions, seems to make this difference. And Moore's use of the indexical "I" in expressing his belief seems to indicate this fact.

Both problems, it seems to me, should raise the question of what it is to *identify oneself as someone*, and why it makes these specific differences to our cognitive lives. My thought is this: From the first-person perspective, my own *actions* and *beliefs* rarely look on a par with other persons' actions and beliefs; my own actions and beliefs appear as something I have to decide, or something that is "up to me." What I want to suggest is that to take a certain person X's actions and beliefs to be *mine*, or to *identify myself as* X, just *is* to take myself to have control over X's actions and beliefs. Perry and I have relevant beliefs and desires in common, but he differs from me in that he alone takes his (Perry's) *actions* to be under his control. Moore and I have the relevant belief in common,

---

explicit statement of this point.

[3]A attributer may achieve the same by using a *quasi-indexical* expression, or a *quasi-indicator*, in Castañeda (1968)'s sense. "Perry believes that *he* (himself) has a class to teach" can also indicate that Perry identifies himself as the person who has a class to teach. Or we may instead use Castañeda's device "*he**" for disambiguation.

but he differs from me in that he alone takes his (Moore's) *beliefs* to be under his control. Perry's problem and Moore's paradox are indeed the same problem at bottom.

The accounts of Moore's paradox along this line have been developed and defended recently, notably by Richard Moran.[4] I don't have much substantial to add to that, although I do want to clarify some of the issues. On the other hand, influential accounts of Perry's problem have tended to go in completely different directions.[5] So our discussion will naturally tilt toward Perry's problem, until we are ready to advance a positive account.

I start from this intuitive but obscure idea of *identifying oneself as someone.* And Perry's problem and Moore's paradox will be reintroduced as subproblems of the problem of self-identification.

1.2 Belief and Self-Identification

According to the traditional conception of beliefs, a person's belief state is completely characterized by *propositions* the person believes. What are propositions? Everyone will agree on this much: A proposition is something that determines a *truth condition*, which can be identified with a function from possible states of the world to the truth values, or equivalently *a set of possible worlds* in which it is true. We may assume that a proposition just *is* a set of possible worlds. The totality of what the person believes can be represented by a set of possible worlds, the intersection of all propositions the person believes. We may call those possible worlds the person's doxastic possibilities or alter-

[4]Moran (1997; 2001).

[5]Among them are Perry's own account (Perry 1979), David Lewis's (Lewis 1979), and Robert Stalnaker's (Stalnaker 1981).

natives. Then the traditional view comes to this intuitive thesis: Specify fully what the world is like according to a person, and then you fully characterize the person's belief state.

This view of propositions and beliefs is not accepted by everyone, of course. However, I don't think this makes our starting point biased. On the contrary, absent a universal agreement, our choice seems ideal, in that it is *minimalist.* Different views of propositions and beliefs differ only in what *more* is needed to properly characterize belief states. Since we will be considering precisely that question, that is, what more is needed to solve Perry's problem and Moore's paradox, it is a good idea to start with this bare-bones picture of propositions. At the end of the day, we might be led to conclude that a proper characterization of a belief state needs to be enriched by something like *modes of presentation.* I am completely open to that, but what I think we need to get clearer is what it is that needs to be explained by such a thing. And I believe that this view of beliefs provides a useful framework for posing and answering such a question.

Some strongly *felt* that this picture of belief left out something important about the location and identity of the first person. Thomas Nagel is one of them:

> Given a complete description of the world from no particular point of view, including all the people in it, one of whom is Thomas Nagel, it seems ... that something has been left out, something absolutely essential remains to be specified, namely which of them I am.[6]

Nagel often says that the *fact* that he is TN is left out in an objective *description* of the world. But I think his point better applies to belief. In the possible worlds Nagel believes might be actual, there will be numerous people. But none of them will be specially marked as the subject of this very belief, which those possible

[6]Nagel (1986), 54. See also Nagel (1983).

worlds are being used to characterize. Take this convenient analogy of a *map*.[7] A map, like a belief state depicted by the traditional conception, purports to represent the world "from no particular point of view." For that reason, the map lacks a resource to represent where the map itself is located. But this information is absolutely necessary for the map to help us find our way around. We should obtain this additional information by aligning the map with what we see in the surroundings. Likewise, doesn't a belief state characterized as a set of possible worlds leave out important information, which is absolutely necessary for the belief to guide actions? However, one may doubt the effectiveness of the analogy. In the case of the map, the additional information is intelligible because we have another representational system to align the map with. But in the belief case, our belief is all there is that we have to go on with. Some might think that this is what makes the problem in the real case more challenging and interesting. Others might think that it makes the whole question unintelligible, and even that the feeling of something's left out is a mere illusion. So there seems to be something that Nagel is getting at, but it is hard to pin down.

There is a simple and familiar philosophical technique that can be usefully applied in this sort of situation: *Imagine two persons who are indiscernible with respect to what they believe, and see whether and how they might differ.* Since we will be considering many such examples, let's give them a label: *two-person cases.* Consider the following case from Stalnaker:

> Suppose O'Leary knows he is in the basement, and that Daniels is in the kitchen. Daniels also knows this—that he is in the kitchen and that O'Leary is in the basement. Each knows who and where he himself is, and who and where the other is. The possible worlds compatible with the two men's beliefs are the same—they don't disagree

[7] See Velleman (1996), 173-8, for an insightful discussion of the map analogy in relation to Nagel's problem.

> about anything, and there is nothing relevant that one believes and the other is ignorant about. Yet there is obviously a significant difference between their doxastic situations: O'Leary identifies himself as the one in the basement, while Daniels identifies himself as the one in the kitchen. This difference between the belief states of the two men is not reflected in a representation of a belief state by a set of possible worlds.[8]

O'Leary and Daniels agree on every bit of what the objective world is like. So according to the traditional view, they are in the same belief state. But intuitively, "there is obviously a significant difference between their doxastic situations." One *identifies himself as O'Leary*, and the other *identifies himself as Daniels*. Stalnaker also says that their difference is "a difference in perspective," rather than "a disagreement." If the existence of such a difference is "obvious" as Stalnaker thinks it is, then the case already establishes that the feeling of something's left out is *not* an illusion.

I think this notion, *identifying oneself as someone*, or *self-identification* for short, obscure as it may be, has some intuitive content, capturing not only what O'Leary and Daniels differ in, but also what Nagel was puzzled at, and what the map analogy purports to illustrate. Moreover, I think it is potentially a very important philosophical notion, which is at the very heart of many issues of the self and first-personal perspectives. In particular, as I will argue, both Perry's problem and Moore's paradox can be regarded as subproblems of the general problem that self-identification raises. And I believe this notion *can* be made precise, as we shall see shortly, with the help of this sort of two-person examples.[9]

---

[8]Stalnaker (1999a), 21. Perry also discusses many two-person examples in Perry (2006).

[9]My notion of self-identification is different from Gareth Evans's, in his marvelous chapter "Self-Identification" in Evans (1982). His notion of "identification" is more theoretically loaded. In his view, identification is involved in reference in general, and is a matter of *concepts*. For all that, what I will say is much congenial to, and in fact influenced by, his view.

Let me first say what I think self-identification is *not*. It is tempting to think that self-identification is simply a matter of believing propositions expressed by identity statements of the form "I=O'Leary" or "I=TN," and that the problem it raises is to explain *informativeness* of such statements. Suppose that Daniels, a moment ago, wasn't sure who was in the basement, although he knew for certain that it was either O'Leary or Fred. Daniels shouts, "Who's there in the basement?" O'Leary's shouting back "I am O'Leary" from the basement may resolve his uncertainty (assuming that Daniels can tell where the voice comes from). So O'Leary's saying "I=O'Leary" is *informative*; that is, it gets across the information Daniels sought, the information the receipt of which Daniels might confirm by saying "You are O'Leary." But it seems that this information *can* be represented within the traditional conception of beliefs. Consider *what the world was like* according to Daniels's belief *before* acquiring the new information: Among the worlds compatible with his belief, some are such that O'Leary is in the basement, and the rest are such that Fred is in the basement. The latter possibilities get eliminated by his acquiring the new information. Presumably, the set of possibilities thus ruled out can be identified with the proposition he learned.

Now it is an important and difficult question *how* "I=O'Leary," made by O'Leary in this context, came to express this proposition. This may look particularly difficult in the face of a widely accepted view that both terms "I" and "O'Leary" are *directly referential*.[10] For if they both directly refer to O'Leary, doesn't "I=O'Leary" just express a necessary proposition that O'Leary=O'Leary? But why think that the way statements made in contexts determine propositions is as simple as that? We may call this problem the *semantic* problem of

[10]Kaplan (1989).

indexical beliefs.[11]

I think the semantic problem of indexical beliefs has to be distinguished from the problem of self-identification. Note that Daniels's acquiring the new information certainly wouldn't make Daniels *identify himself with* O'Leary. So O'Leary's identifying himself as O'Leary is not captured by the information conveyed by "I=O'Leary."

1.3 Self-Identification and Perry's Problem

Then what *is* identifying oneself as someone? The approach I will take and promote is an *externalist* and *functionalist* one. As Nagel broaches the issue, he says: "I shall speak about the subject in the first person, in the Cartesian style which is intended to be understood by others as applying in the first person to themselves."[12] This sort of approach is quite typical in dealing with various problems concerning first-personal perspectives. But if one thinks that the issues about the first-personal perspectives *ought to* be approached from the first-personal perspective, it is simply to confuse the objects of study with the methods of study. I proceed differently: I shall rarely speak about the subject in the first person, but *in the third person.* Instead of asking how I am "presented to myself" *within*, I ask what *external* differences self-identification makes to our cognitive lives. We may suppose that self-identification is not there for nothing, but to play distinctive *roles* in our rational activities. The benefit of bringing out *two-person* examples is that it makes it easier to identify

[11] My thought on the semantic matter was much influenced by Stalnaker (1981), in particular, by his "holistic approach" on the semantic problem (137). In that paper, Stalnaker thought that the semantic problem of indexical beliefs could resolve all problems, including Perry's problem, but later concedes that there is some residual problem (Stalnaker 1999a, 19-21). The example of O'Leary and Daniels we quoted earlier occurs in that context.

[12] Nagel (1986), 55.

such roles. By considering what *differences* there may be between two persons sharing all beliefs, we may be able to identify the roles that should be played by self-identification. Then we can take self-identification to *be* the mental feature that plays such roles.

There may be various psychological differences between two persons sharing beliefs, but some will be simply irrelevant to self-identification; e.g., their desires, their characters, and their intelligence, etc. So this approach needs to be guided by some intuitive ideas we have about the notion of self-identification. Here are some: First, since it is identifying oneself as a particular person in the world as the subject takes it to be, it will have to be something that determines a "center" in each of the subject's doxastic worlds. Second, self-identification should be some *doxastic* feature, not anything like pro and con attitudes. (That's why we say that the traditional conception of *belief* leaves something out.) The crucial constraint this imposes, I think, is this: Self-identification should be something that a person in principle can get *right* or *wrong*. In other words, it should be something that makes self-*mis*identification intelligible. Third, it should be something relevant to *rationality*. The rationale for this is the idea of *constitutive rationality*, broadly construed, by which I mean the principle that mental features in general should be understood by their roles in rational activities. This assumption is what grounds our functionalist approach. We don't want self-identification to be a matter of idle feelings that don't make any substantial difference to our cognitive lives. Or at any rate, if it turned out to be so, I would conclude that nothing was left out in the traditional conception of beliefs after all.

So what relevant differences might there be between two persons sharing beliefs? I think two kinds of differences are particularly important:

First, from the "output" side: Suppose that TN in New York and JP in Stanford share all their relevant beliefs. JP has a class to teach shortly, and that both JP and TN not only know this, but also want JP not to miss the class. But it is only JP that gets up and start moving for the campus. What is a rational thing for TN to do, if they are talking over the phone, might be to urge JP to hang up and go. That is, TN and JP apparently *share all relevant beliefs and desires*, but will be *disposed to act differently*, without anyone failing to be rational thereby. This must be a significant difference relevant to self-identification. Note that what we called *Perry's problem* is precisely the problem of explaining this difference. So Perry's problem, as I understand it, is the problem of accounting for the "output" component of self-identification.

Second, from the "input" side: Suppose that a gas leak in the house caused O'Leary and Daniels to pass out and to be taken to an emergency room. They both wake up, without knowing where they are, but still sharing all beliefs. Both look around, and see a characteristic scene as of an emergency room. What O'Leary immediately learns from his experience is that *he* (that is, O'Leary) is in a hospital, while what Daniels learns from his experience is that *he* (that is, Daniels) is in a hospital. They started with apparently the same beliefs, but upon having apparently the same experience, they come to have *different* beliefs.[13] Again, this seems to be a significant difference relevant to self-identification.

[13]I am aware that there are some tricky issues about taxonomy of experiences involved in stating this precisely. Suppose that two experiences are of the same type if and only if they don't differ in their contents and phenomenology (not excluding the possibility that one determines the other). Under this condition, should not O'Leary and Daniels's experiences be different? Their beds might be placed side by side, but still, the way things look may be slightly different due to the different angles from which they see things. To perfectly match our needs, we will have to consider experiences other than visual kinds. Suppose that both O'Leary and Daniels hear loud sirens, but in such a way that they can't locate the source of the sound. O'Leary will be disposed to believe something about O'Leary's location, while Daniels about Daniels's location.

It seems intuitively right that these two components, the *output* and the *input* components, play central roles in self-identification.[14] We may want to remind ourselves of Daniel Dennett's vivid fantasy in this context.[15] Dennett's brain is taken out of his skull and kept in a vat in a laboratory in Houston, TX, and is remotely controlling his body, which is undertaking a dangerous subterranean mission in Tulsa, OK. His brain gains information from his body, and controls his body, in such a way that he would not have noticed the difference if he hadn't been informed of his extraordinary situation in some other way. Dennett asks himself, "Where am I?" He finds it almost impossible to identify himself as the brain. He reports, "I tried and tried to think myself into the vat, but no avail."[16] It is only when all connections from and to his body are severed that Dennett is forced to "think himself into the vat." It seems to me that this kind of thought experiments shows how tightly our intuitive notion of self-identification is related to perceptual inputs and behavioral outputs. Put in somewhat memorable phrases, one's identifying oneself with TN seems to be a matter of regarding himself as *acting through TN's body* and as *seeing through TN's eyes.*

(But doesn't the fact that Dennett *could* identify himself with his brain in the vat, *after* the body was destroyed and all connections from and out of his body were severed, imply that there are some other components of self-identification,

---

[14]See Evans (1982) for a similar observation. But Evans says this, referring to Perry's and Lewis's work: "Neglect, in this work, of the other element produces a strangely one-sided effect, 'strangely,' because the other element is just as striking, and clearly parallel, and also because the dominant conception of the identification of empirical content concentrates exclusively on the *input* or *evidential* side of things. This chapter will partly redress the balance by rather neglecting the action component." (Evans 1982, 207, fn. 4). I am inclined to disagree with him on two things: First, I don't think the input side is dominant over the output side. Second, I think that we need to "redress the balance" in the opposite direction.

[15]Dennett (1981b).

[16]Dennett (1981b), 312.

distinguished from the input and output components? Two points: First, the sense still remains that the behavioral output and the perceptual inputs are dominant sources of self-identification. But, second, I am inclined to think that we don't need to posit a wholly distinct component of self-identification to accommodate such cases. For inputs and outputs don't have to be *bodily*, that is, perceptual and behavioral. We can make sense of *mental* inputs and outputs too. What I will be saying about Moore's paradox later can be seen in this light.)

I am not entirely comfortable with speaking as if there were distinct "components" of self-identification. For is it conceivable that one regards oneself as acting through TN's body, but at the same time, as seeing through JP's eyes? If not, wouldn't it be because the two components of self-identification are necessarily connected? Maybe, or maybe not. However, methodologically, the best strategy seems to proceed as if they could be separately dealt with, only to make things manageable. I will focus on *Perry's problem*, that is, the problem of accounting for the output component of self-identification, in the next few sections.[17]

I should emphasize again that our approach to self-identification and Perry's problem is a *functionalist* one. I think we can make some progress on this difficult subject by, and only by, taking such an approach. This approach naturally imposes some strict standard that various accounts should meet. If an account of Perry's problem locates or posits a difference between O'Leary and Daniels, but that difference is not something that is up to the *role* it is supposed to play,

[17]I believe that Perry himself won't take strong issue with this formulation of Perry's problem. But there is an important difference between his way of seeing the problem and mine. He draws a line between the semantic problem and the problem of self-identification differently than I do, in such a way to give the semantic part a less burden. For example, suppose that Perry knows that Perry is making a mess in a supermarket, forgetting for a moment his identity. When Perry realizes that *he* is making a mess, there is some change in his belief state. But Perry would not characterize this change in terms of acquiring new information.

then we should be suspicious of the account. In the following two sections, I will consider some specimen accounts, and argue that they fail to meet this standard.

1.4 Limited Accessibility

We claimed that the semantic problem of explaining informativeness of identity statements such as "I=TN" should be distinguished from the problem of self-identification. I was guided by what I take to be an obvious principle: A statement is informative *only* to the extent that it can get across information to other people. But couldn't there be some information *privy to O'Leary*, which is *not communicated* when he says "I=O'Leary," but nonetheless is part of *what* he believes? This leads to what Perry calls *limited accessibility.*[18] The idea is that there might be a proposition that JP believes but TN doesn't (because he can't), and that this difference might explain their different dispositions to act.

The idea that certain propositions are accessible only to some persons in certain limited contexts or situations by itself, I think, is neither incoherent nor mysterious. In fact, I think any reasonable account of the *semantic* problem of indexical beliefs should be able to make sense of this kind of limited accessibility. Suppose that I meet a person at a philosophy conference, who for all I know might be either TN or JP. The person's saying, "I am TN," will resolve my uncertainty; that is, it lets me eliminate possibilities in which that particular person that I am facing is JP. However, when I get home and try to tell my wife what I learned from that conversation, I will say something like "I met TN in person at the conference." But that's not exactly the information I learned from TN at

[18]Perry (1979), 37.

that moment: it was information about that particular person, TN, not directly about myself. It seems that my wife can't grasp the same proposition, because she cannot distinguish between certain possibilities I can distinguish. On the other hand, a friend of mine who was there with me won't have any problem with grasping that proposition. He and I may chat about what happened at the conference, and we may say things like "That person was TN." It seems that being in a position to demonstratively refer to a particular person or thing, by virtue of being adequately related to the person or the thing by experience or memory, seems to make difference to what possibilities one can distinguish.[19]

Then could we push this point a little further, and use this feature of indexical beliefs to locate O'Leary and Daniels, or TN and JP? If one is in the basement, and the other is in the kitchen, then isn't there a sense in which they are in different contexts or situations? But their physical distance is only an accidental feature of the example. It seems that we can readily make the two persons exactly in the same situation or context, with all the points of the example intact. O'Leary and Daniels may finally meet in the dining room, sitting across the table, still sharing all beliefs. Each should be in a position to demonstratively refer to the other, and any other features in the surroundings. That is, there seems to be nothing to which only one person is in a position to demonstratively refer to. But the difference in self-identification, of course, should still remain.

In order for the idea of limited accessibility to be of help, we need to push the "context or situation" inside one's "internal world." There seems to be something inside O'Leary's internal world that only he can be in a position to demonstratively refer to; *his own token thoughts and experiences.* Then that O'Leary

[19]Here I am following Stalnaker (2008). But he seems to think that this consideration can be extended to explain self-identification. But I don't think this works, as I will argue shortly.

is having this particular experience may be the proposition that only he can have access to. But what if Daniels is a neurosurgeon who is directly observing O'Leary's brain? If you are a physicalist, shouldn't you say that Daniels can also believe the same proposition—something he would express by saying, "You are now thinking *that* thought," pointing to a certain brain activity? If you think Daniels can't *in principle* be in a position to refer demonstratively to O'Leary's token experience or thought, then you are committed to the view that there exists something in his "private realm," such as *sense-data*, to which only the subject can possibly be in a position to demonstratively refer to.

In this way, we may be able to make room in the space of possibilities for the difference between their beliefs. But obvious metaphysical worries aside, I don't see how this can be a *relevant* difference. Although Perry explicitly distances himself from such a view, he seems to think that such a view *could* provide an adequate answer to Perry's problem. He writes:

> Such a belief [the belief that *the person with this sense-datum* (attending to one of my sense-data) *has a class to teach*] not only makes it well-advised to execute the movements necessary to get oneself to class, it can motivate it, since everyone who believes that proposition is well-advised to execute those movements.[20]

Why *only* JP, *not* TN, starts moving for campus, despite their apparently sharing all relevant beliefs and desires? Perry's suggestion is that JP's believing something he would express, referring to his experience, with "The person with *this* sense-datum has a class to teach," can explain that. TN does not move to actualize this, because he does not, or rather *cannot*, have this belief.

However, many will probably share with me the sense that there is something deeply unsatisfactory with this supposed explanation. Perhaps we can

[20]Perry (2006), 215.

admit that there *is* a sense in which such an explanation works. Perry cites as a reason the fact that "everyone who believes that proposition is well-advised to execute those movements." Presumably, this universal generalization is not only true, but also "law-like" (although this is so only because the proposition is so chosen as to guarantee that only JP has access to it). And, of course, whenever we have a true law-like universal generalization, we can explain an instance of it, by subsuming it under the generalization. But what we want in this context is not (just) *causal* explanation, but *rationalizing* explanation. The existence of such a universal generalization may ground a causal explanation, but by itself doesn't guarantee that it is a rationalizing explanation. What we want is an account of why a rational person ought to fall under that universal generalization.

As far as I know, no one explicitly endorsed the sense-datum account for Perry's problem. But there are many similar approaches that have attracted philosophers. For example, Frege notoriously says, "everyone is presented to himself in a special and primitive way, in which he is presented to no one else."[21] And he says that the subject alone "can grasp thoughts specified in this way." I think that the same objections apply to this. First, it is highly doubtful whether even Frege can avoid unpalatable metaphysical consequences.[22] But the real worry apart from that is that this extravagance doesn't seem to buy anything, as far as Perry's problem is concerned.

[21]Frege (1919/1997). 333.
[22]For this point, see Perry (1977), 15 and Stalnaker (2003b), 261, fn. 10.

1.5 Lewis's *De Se* Content

David Lewis's influential account goes in a different direction.[23] Lewis proposes to revise our conception of *objects* of beliefs (and other attitudes) systematically. Instead of propositions, he takes *properties*, or equivalently (in his framework) sets of *centered possible worlds*, to be objects of beliefs. A person's beliefs are characterized in terms of properties the person *self-ascribes*. For example, when O'Leary says "I am O'Leary," he is self-ascribing the property of being O'Leary. Even thoughts apparently not about oneself can be reinterpreted in this way. For example, believing that Earth is round amounts to self-ascribing the property of inhabiting a possible world wherein Earth is round. We can call those properties or sets of centered possible worlds self-ascribed *de se* contents.

We distinguished a couple of different problems in the vicinity of Perry's problem. There was the *semantic* problem explaining informativeness of identity statements, such as "I am O'Leary," made in certain contexts. We claimed that the problem of self-identification should be distinguished from that, and separated out two different components of it. As I understand it, Lewis's account is designed to give a sweeping solution to all those problems at once. Some might regard this *unifying* feature to be a merit of his account, but I am not so sure. For, if they (the semantic problem and the problem of self-identification) are really *conceptually distinct* problems as I think they are, then treating them as if they were one might be not so much a *unification* as a *conflation*.

I suspect that some distorting effect of his way of dealing with the semantic problem is indeed symptomatic of a conflation. O'Leary's saying "I am O'Leary,"

[23] Lewis (1979; 1994).

resolved Daniels's curiosity about who's in the basement. How did we explain what happened? Simple: O'Leary said what he believed, and Daniels came to believe what O'Leary said. But how could Lewis's account explain this? The explanation would start with O'Leary's self-ascribing *being O'Leary*, and end with Daniels's self-ascribing *being facing a person who is O'Leary.*[24] But how to fill out the middle steps in a general way is quite unclear. This may not be impossible, but I don't see much point in making things so convoluted. Lewis did see some point in it, of course, because he believed his account could solve some other (in my view, conceptually distinct) problems. So let's see how well it does on these.

It seems that Lewis's account addresses the input component of self-identification nicely—and I suspect this is responsible for much popularity it enjoys. O'Leary wakes up in an emergency room, having no idea where he is. Looking around and seeing a familiar scene of those white colors, he realizes that *he* is in a hospital. How would this be possible if his experience did not already contain the information *about himself*? Lewis's account explains this by assuming that his experience already has a *de se* content *being in a such and such room.* O'Leary simply self-ascribes the *de se* content he picks up from his experience.[25]

(Although we can't go into this issue, I should mention that Lewis's position on this matter is *not* indisputable. It is true that we most often acquire information about our location and identity from what we experience. But it does not follow from this that perceptual information itself is already "imbued with subjective significance" (to parody a phrase from Evans[26]). One reason for doubting this is basically a Humean one: Carefully attending to my experience,

[24]This and other related objections were forcefully raised by Stalnaker (2008), 50-1.
[25]Lewis (1979), 138-9.
[26]Evans (1982), 123.

I just don't find myself at all represented there, even implicitly (whatever this could mean). I think a better alternative is this: Experiences themselves have only objective contents, but I am disposed to form beliefs about a particular person on the basis of it. Such a disposition is constitutive of my identifying myself with that person.[27])

Our main concern, of course, is with the output component of self-identification, that is, Perry's problem. Lewis's account seems to have achieved the reputation that it does equally well here. I'll first say where this impression may come from, and why I think it is a mere illusion.

De se contents resemble *narrow contents* in many respects. Lewis himself seems to think that his view is supported by the same reasons that support narrow contents. Consider another familiar figure, mad Heimson, who madly thinks he is Hume. We can contrive the case a little further, by supposing that Heimson "got his head into perfect match with Hume's in every way that is at all relevant to what he believes." Heimson believes (wrongly) that *he* is Hume, while Hume believes (correctly) that *he* is Hume. Under any attempt to characterize the two persons' relevant beliefs in terms of propositions, they will *differ* in *what* they believe, despite their perfect match: Heimson believes of himself (that is, Heimson) that he is Hume, while Hume believes of himself (that is, Hume) that he is Hume. Lewis writes:

> If ... Heimson and Hume do not believe alike, then *beliefs ain't in the head!* They depend partly on something else, so that if your head is in a certain state and you're Hume you believe one thing, but if your head is in that same state and you're Heimson you believe something else. Not good.[28]

[27]This is basically Evans's view as I understand it. See Evans (1982), Sec. 7.4. As we shall see, the account of Perry's problem I propose is quite congenial to this treatment of the input component.

[28]Lewis (1979), 142. See also Lewis (1994).

However, according to Lewis, Heimson and Hume share the *de se* content *being Hume*, and this *de se* content is "in their heads." (One thing to notice is that although it is true that they share the property of self-ascribing *being Hume*, it doesn't follow from this that this property is "in their heads" or intrinsic to them. The latter is in fact very suspicious, for the familiar externalist reason. But let it pass for the sake of argument.)

This reasoning leading to *de se* contents is quite similar to that leading to narrow contents. Consider another familiar story: Oscar on Earth and his molecule-by-molecule duplicate on Twin Earth, Twoscar, have different thoughts when they are looking at lakes from their respective places, one thinking that there is *water* in the lake, the other that there is *twater* in the lake. "Their beliefs ain't in the head! Not good," proponents of narrow contents would say. To get at the beliefs in the head, we'd better "factor out" the content in the head (that is, narrow content) that Oscar and Twoscar share.

But what's so bad about beliefs' *not* being in the head? Why do they think that having certain beliefs must be intrinsic properties? One major motivation has much to do with *action explanation.* The friends of narrow contents often appeal to the principle that mental states should be individuated with respect to *causal powers*, for them to be *explanatorily* relevant.[29] Being duplicates, Oscar and Twoscar will behave in the same manner; for example, they both would be reaching for glasses of water and twater, respectively, to quench their thirst. So Oscar's belief and Twoscar's belief must have a common causal power, and in order to individuate their beliefs with respect to causal powers, we need to posit narrow contents their beliefs have in common. Likewise, for Hume and

[29]Lewis himself, however, explicitly refuses to appeal to such a principle. See Lewis (1994), 316.

Heimson, who are intrinsic duplicates, *de se* contents that are shared by them must be more apt as objects of belief for *action explanation.* Why does only JP, but not TN, who has the same beliefs and desires, move for the campus? It is because only JP self-ascribes the property of *having a class to teach.* Another person, say DL, who shares the *de se* content with JP, would be disposed to act in a similar way as JP does.

I suspect this line of thought might have tempted many to think that Lewis's account can provide an adequate answer to Perry's problem. But I think this is a mistake. Whatever problems narrow contents are designed to solve, the issue of "being in the head" seems irrelevant to *our* problem at hand, that is, Perry's problem. The complaint that was raised against contents not in the head was that they, being intrinsic properties *plus* external relations, contain *more* than necessary and relevant to action explanation, and so fail to isolate causally relevant properties. So they are incapable of explaining the similar behaviors of Oscar and Twoscar. We need to *factor out internal contributions* of what Oscar and Twoscar (widely) believe. But our problem is quite different: The case of JP and TN is problematic because their propositional beliefs, wide or not, together with desires, seem to *underdetermine* their actions. The question is what *more* is needed, besides propositional beliefs and desires, to account for their difference in actions. This *additional* something can't be gotten by such a method of "factoring out," by simple arithmetic reasoning, so to speak.

But, one might protest, isn't "everyone who self-ascribes *having a class to teach,* other conditions being equal, is motivated to act to realize the property of *leaving for campus*" true, and so doesn't it explain why only JP, not TN is leaving for campus? What this generalization says in effect is that everyone who is in a position to say "I have a class to teach" is disposed to leave for

campus. But this is just an observation of phenomena to be explained, not itself an explanation. Re-classifying beliefs in terms of *de se* contents may help, in that it sharpens what needs to be explained. But such a classification should be only the beginning of the inquiry, not the end of it.

1.6 An Agent's Options and States of the World

The two accounts we considered in the two previous sections both attempt to locate the ground of self-identification in *what* a person believes. There is a perfectly sensible motivation behind this sort of move. Consider the following line of reasoning. We have a fundamental principle of folk-psychology: *What an agent believes, together with what he desires, determines what he is rationally to do.* This principle is supposed to be *a priori* and analytic, partially defining the concept of *rationality* itself. So we'd better hold on to this principle. Moreover, *what* an agent believes and desires should be determined by the role they play in rational explanation, the form of which is defined by this principle. Therefore, in the face of apparent counterexamples to the principle, such as the case of JP and TN, the only reasonable way to go is to conclude that *what* JP and TN believe are not the same after all.

There is much to respect in this line of reasoning. Its only, but crucial, defect is that it gets "the fundamental principle of folk-psychology" too simplistic. We need to get clearer about the structure of folk-psychology and action explanation. I think looking at *Bayesian decision theory* will be helpful for that; for it is supposed to be the most sophisticated refinement of folk-psychology.

In Bayesian decision theory, an agent's decision problem is modeled by a space of possible worlds, as finely grained as relevant to the given problem. The

space of possibilities is supposed to represent the agent's *doxastic alternatives.* So the space represents the agent's whole belief, according to the traditional view of belief. Bayesian decision theory, being Bayesian, also assumes that an agent's belief about the world comes in a more fine-grained form, determining a *probability* distribution over the space of doxastic possibilities, but nothing important will hang on this additional complexity. Each possible world is associated with a *value*, representing the extent to which the agent wants it to be actualized. It is assumed that the value of a proposition, or a set of possible worlds, can be determined as the average of those values assigned to those individual possible worlds, weighted by probabilities assigned to them.

To see this theory in action, let us embellish the example of O'Leary and Daniels a little bit. Suppose that they both want to divide household chores evenly and effectively, and that they both know that the best way is for each to do his job in his respective place: For O'Leary, it is doing laundry in the basement, and for Daniels to make the omelet in the kitchen. As before, they are assumed to share all beliefs (including their probability distributions) and values. We can represent their decision problems in the following *single* matrix, thanks to these assumptions.

		D	
		Cook Omelet	Not Cook
O	Do Laundry	4	1
	No Laundry	1	0

The space represents their doxastic alternatives. Each cell represents a possible

consequence of their actions: Within each cell, they both don't care about which possibility is actualized.

So all these are common grounds for O'Leary and Daniels. But decision theory does *not* see their decision problems as the same. For it recognizes *additional* determinants of their decision problems: It distinguishes between *what each agent has control over*, and *what holds independently of what the agent does*. Leonard Savage, in his seminal work, conceptualized these two ideas as *acts* and *states*, and later David Lewis refines these notions further and gives them different names: *options* and *dependency hypotheses*, respectively.[30] I will follow Lewis's formulation, although I will use each pair of notions interchangeably ("states" sounds more standard than Lewis's idiosyncratic "dependency hypotheses," but the latter seems more descriptively correct). Savage identifies acts with a function from states to consequences, but Lewis takes an individual act to be a proposition. So in our example, that O'Leary will do laundry identifies an act or option for O'Leary. His not doing laundry is not to refrain from acting, and it forms another option for him. So options in a given decision situation forms a *partition* of the space of the agent's doxastic possibilities. More precisely, Lewis says:

> Suppose we have a partition of propositions that distinguish worlds where the agent acts differently .... Further, he can act at will so as to make any one of these propositions hold, but he cannot act at will so as to make any proposition hold that implies but is not implied by (is properly included in) a proposition in the partition. The partition gives the most detailed specifications of his present

[30]See Savage (1972), Ch. 2 and Lewis (1981). I am developing the basic idea in terms of *causal* decision theory, partly because, like many, I think that's the correct decision theory, but also partly because the contrast between acts and states is particularly relevant to what I will say. But it is important to note that the notion of *acts* is not something even *epistemic* decision theory can dispense with. Without it, epistemic decision theory will be at best "a theory of preferences," not a *decision* theory. See Jeffrey (1990), 83-4.

> action over which he has control. Then this is the partition of the agents' alternative *options*.[31]

On the other hand, a state or dependency hypothesis is understood as a proposition about what the world is like in relevant respects beyond the agent's control. More precisely, it is "a maximally specific proposition about how the things he cares about do and do not depend causally on his present actions."[32] Assuming that each state is as "maximally specific" as relevant, then states also effect a partition of the space of the agent's doxastic possibilities. Over what state or dependency hypothesis holds, the agent has no control; that is, he can't make one or another dependency hypothesis true.

Although O'Leary and Daniels share their beliefs and desires, there is a sense in which their decision situations are *diametrically opposite* in this particular setup: One person's options are the other's states, and *vice versa*. O'Leary's two options are doing laundry and not doing laundry; the two propositions that he can act to make true. He does not know for sure which consequence each of these options will yield, because he does not know for sure what Daniels will do. The two alternatives about what Daniels will do are states for O'Leary. But precisely the same alternatives are *acts* for Daniels. Unlike O'Leary, Daniels has control over them; he can act to make one of the two alternatives true.

Decision theory prescribes an agent to behave in a way to actualize the option with the higher *expected utility*. The expected utility of each option is calculated as the sum of values of possible consequences, weighted by probability assigned to each *state*. The agent may assign some probabilities to his own options, but what probabilities he assigns across his own options is simply irrelevant, that is, doesn't appear at all in the calculation. So that's how decision

[31]Lewis (1981), 308.
[32]Lewis (1981), 313.

theory prescribes O'Leary and Daniels to act to bring about different possibilities, even if they share beliefs and desires. So seen in the context of Bayesian decision theory, Perry's problem does *not* arise.[33]

Let's now ask a foundational question: how are each person's options determined? If O'Leary and Daniels share all beliefs and desires, then how can their decision situations be diametrically opposite? This may look like a trivial question, but it turns out to be much more difficult to answer than it may seem at first.

One might think that the agent's options are determined by some *facts* about the agent and the circumstances, rather than a certain *doxastic* feature attributable to the agent; in our case, the facts about what O'Leary and Daniels *can*, or have power to, do. This won't do, however. The agent's options partition the space of the agent's *doxastic* alternatives, and of course, the agent may get facts wrong. Suppose that as a matter of fact, Daniels *can't* make the omelet, because the eggs are all rotten, or because he is going to pass out very soon due to an impending gas leak (all unbeknown to him at the moment of decision). Or what if as a matter of fact, the universe is *deterministic*, so that strictly speaking, nothing is under Daniels's control? But none of these seems to affect what options he *takes himself to be confronted with*, and that's what is relevant, rather than what options he actually has power to actualize.

[33]It is interesting and encouraging that Perry himself, in his recent paper, observes something quite similar. As he summarizes the lesson he learns from his examples, he says: "Consider any person x who wants to achieve some goal G that requires changes in the world outside of his own thought. There will be a set of bodily movements that person can execute that will bring about results that will promote G. For most of us and most of the desires or wishes we might have, and many of our goals, the set will be empty. There is nothing I can do to prevent Hitler from invading Poland or even make it less likely that he did so. ...But there are things I can do to, say, help the hurricane victims in Haiti, or bring it about that this article get finished. Let's say that the set of movements that would promote my goals are ones I would be *well advised* to make" (Perry 2006, 214). Here what Perry means by "a set of bodily movements that person can execute" amounts to *acts* in our sense.

Then an obvious thought will be that Daniels's options are determined by his *belief* on this matter; that is, Daniels believes that his making the omelet or not is under his control. But this by itself won't do either. For obviously O'Leary easily *can* believe exactly the same thing, that is, that Daniels's making the omelet or not is under Daniels's control. Yet, of course, O'Leary's believing it won't make them *his* options. One may want to insist that Daniels's believing that *he himself* has control over those possibilities makes difference. But this just begs the question in the present context. So we are led back to the idea that there must be some doxastic feature that is not captured by the characterization of an agent's belief in terms of propositions, but that is substantially employed in decision theory.

Recall that we started out with the idea that we might be able to define self-identification by its *functional roles.* Two-person examples convinced us that there must be some distinctive role played by it in rational actions, but fell short of pinning it down. Now if we use decision theory as a guide as I think we should, we come to have a more articulate vision of what it is: it should be something that can *determine a partition within the space of one's doxastic alternatives*, the intuitive interpretation of which is that the agent *takes himself to have control over* which partition to be actualized. Then the doxastic feature that is supposed to be left out in the traditional conception of belief must be something that could play this role.

That the missing feature is best represented as a way of *partitioning* the space of doxastic possibilities, I think, has a great theoretical significance. We can now see a more general reason for why various attempts to locate the ground of self-identification in what a person believes, or the space of the person's doxastic alternatives, are doomed to failure. Modify the space of O'Leary's

doxastic possibilities in whatever way you like. You may throw in more possibilities that are supposed to be distinguishable only by O'Leary (as in the limited accessibility view). Or you may overhaul the space by replacing possible worlds with *centered* possible worlds, all centered on O'Leary (as in Lewis's account).[34] But however you modify the space of possibilities, you simply cannot read off from it how an option *partition* is drawn inside the space. Then such a move of modifying the space of possibilities looks rather pointless. And that's also the reason why I think the semantic problem is *conceptually distinct* from Perry's problem: If the former concerns the whole space of an agent's possibilities, the latter concerns how to partition it.

We have considered Perry's problem within the decision theoretic framework. I think that formal decision theory is descriptive rather than revisionary, meaning that it reveals the structure of folk-psychology that was already implicit in our practice. I suspect that failing to get clearer about the structure of folk-psychology is responsible for many misguided attempts to solve Perry's problem. Of course, a substantial and difficult question still remains what these different ways of partitioning the space of possibilities amount to in the real thing. Before considering that, however, I finally turn to Moore's paradox; for the account I will be proposing later draws on an account that has been suggested for Moore's paradox.

[34]As Lewis well notices, replacing possible worlds with centered possible worlds does not affect the basic structure of decision theory. He says: "It is interesting to ask what happens to decision theory if we take all attitudes as *de se*. Answer: very little. We replace the space of worlds by the space of centered worlds, or by the space of all inhabitants of worlds. All else is just as before" (Lewis 1979, 149).

1.7 Moore's Paradox and Self-Identification

Moore's paradox is most often formulated as a problem about some peculiar *linguistic* phenomena: *Asserting* the sentences of the forms, "p, but I don't believe p" or "I believe p, but p is not true" (which we may call *Moore-paradoxical sentences*), sounds absurd, or even like a flat-out contradiction. But the problem is that what is being asserted, that is, the proposition expressed by such an assertion, cannot be seen as contradictory. We can clearly envision the situation in which my assertion of "It is raining, but I don't believe it," turns out to be true. A moment later, I may say: "It was raining, but I didn't believe that a moment ago." But asserting this is in no way absurd or contradictory. However, presumably, what I said earlier is the same as what I say now. So it is paradoxical.

By far the most popular approach to Moore's paradox is to see it as some sort of pragmatic impropriety. Many of the pragmatic approaches appeal to *the nature of assertion.* As I assert "it is raining," I *imply* that I believe or know this, or I am *representing myself as knowing or believing* it.[35] Suppose I assert, "It is raining, and I don't believe it." Then by asserting "it is raining," I am implicitly representing myself as believing that it is raining. But the second conjunct explicitly denies it; hence, asserting the conjunction should be "contradictory." Perhaps there is something intuitively right with this approach, but I don't find it, as it is, theoretically satisfactory. Assertions are acts, and the account assumes that when performing a certain sort of acts, one represents oneself as being in a certain way. Does every sort of acts represent the actor as being in a certain way? Or is it only true of linguistic acts? Do we in general need to posit

[35]For the former see Moore (1944/1993), and for the latter see, e.g. DeRose (1991).

representational contents of acts to account for other sorts of linguistic phenomena? Perhaps these questions are not impossible to answer, but it needs to be embedded within a broader theoretical context.

But a much more serious problem with the pragmatic approach along this line is this: By locating the source of the problem in the nature of acts of *assertion*, it makes it difficult to explain why *believing* or *judging* something expressed by those sentences is no less problematic.[36]

One should not jump to the conclusion that there is something wrong with the pragmatic approach *per se*. On the contrary, I think it is the only reasonable option for the linguistic puzzle. That asserting Moore-paradoxical sentences sounds contradictory is a *datum* to be explained. If semantics can't explain it, I don't know of any other alternative than pragmatics. The proper lesson to draw is rather about a methodological one. We'd better start from the mental side. And when we have some reasonable account in hand, we may be able to incorporate that account into the semantic or pragmatic approach.

So what is wrong with a person who believes something he would express with a Moore-paradoxical sentence? It should be clear by now that it cannot be found in what the person believes; for he may well be right on that, and a different person can believe the same thing without any problem. So the lesson we should learn from Moore's paradox is also that there is some doxastic feature that is left out in the traditional conception of beliefs.

And I think this missing feature has to be related to *self-identification* in our sense.[37] Recall how we've been approaching the problem of self-identification:

[36]One may attempt to interpret judging as an internalized version of acts of asserting. Shoemaker (1995)'s account is a sophisticated form of this approach.

[37]Another philosopher who sees the connection between Moore's paradox (although not under that label) and self-identification is Evans (1982). Recall his famous passage (225) about "transparency of beliefs" appears in the chapter titled "Self-Identification."

it is something responsible for relevant doxastic differences between two persons who agree on every bit of what the objective world is like. And I think Moore's paradox points to a significant difference relevant to self-identification. Consider again our O'Leary and Daniels. The moment of decision for O'Leary and Daniels has passed, and they are now sitting across the table. Daniels asks O'Leary, "Did you do laundry?", and O'Leary answers, "Yes, I did." Daniels casually takes his word for it, and comes to believe that O'Leary did laundry. Up to this point, we can assume, they agree on everything propositional. But suppose that they both know that O'Leary hasn't been very credible on this sort of matters, although the track record isn't decisive. O'Leary starts to suspect that Daniels might doubt him because of that, and eventually concludes that Daniels doesn't believe that he did laundry. That conclusion is not something forced upon him by given evidence (which they share), but, we may assume, it is nonetheless a permissible one. O'Leary is now in a position to say, "I did laundry, but Daniels doesn't believe it." However, Daniels himself cannot bring himself to believe that, for fear of Moore's paradox. For if he did, he would be in a position to say "O'Leary did laundry, but I don't believe it." That Daniels doesn't believe that O'Leary did laundry is something permissible to believe from O'Leary's point of view, but not from Daniels's point of view. This must be a relevant difference for self-identification. By our standard then, whatever is responsible for this difference, we should take to be constitutive of self-identification.

Would we then need to posit yet another "component" of self-identification, along with *acting through X's body* and *seeing through X's eyes*, perhaps something that can be phrased in a parallel fashion, "*believing through X's mind*"? Fortunately, however, I think we can avoid this extravagance, by giving a unified

account of *acting through X's body* and *believing through X's mind.* A motivating idea is this: Actions and beliefs are two paradigmatic cases to which the notion of *rationality* finds most natural applications. Just as we take ourselves to have rational control over our own actions, so we take ourselves to have rational control over our own beliefs.

1.8 Self-Identification and a "Deliberative Stance"

In this section, I want to propose an account of the missing feature that is supposed to be left out in the traditional conception of beliefs. I will first explain what this feature is as clearly as I can, and then argue that the feature has the right properties to ground self-identification (or more precisely, the output component of it). The account I will present draws on Richard Moran's work on self-knowledge of *beliefs* (and also Moore's paradox)[38], and will be discussed in a greater detail in Chapter 3, "What Is the First Person Perspective."

Moran contrasts two different *stances* one takes toward certain states of affairs (or equivalently, toward certain *propositions*). One most naturally takes a *theoretical stance* toward other persons' beliefs and actions, or any other states of affairs not involving oneself for that matter. But toward *one's own beliefs and actions*, one almost always takes (or according to Moran, *ought to* take) a *deliberative stance.* The intuitive idea that Moran intends to capture with this notion of a deliberative stance, I believe, is precisely the idea that we have expressed with "taking oneself to have control over" certain states of affairs. Moran often says that taking a deliberative stance toward those states of affairs is a matter of taking those to be "up to me," or "making up mind" on those

[38]Moran (1997; 2001).

matters.

What *is* it to take a certain stance toward some propositions? Moran, in effect, offers an *epistemological* reading. He says, "We should ... see [the deliberative stance] and [the theoretical stance] as *two ways of coming to know the same thing.*"[39] Rather than directly speaking of knowledge, I want to propose the following definition: *A stance is something that determines a pattern of what kinds of considerations (that is, propositions) are* relevant *to resolve a given question.*[40] Perhaps a stance in this sense can be thought to consist of certain *assumptions* plus *epistemic rules*, but details don't matter. For instance, we may speak of taking a "biological" stance toward questions about someone's behavior. This means that we take some propositions about, say, the genetic traits of the person to be relevant to resolve the question about the behavior. Or we may take a "sociological" stance toward exactly the same question, if we take considerations such as the person's class to be relevant. A *theoretical* stance then can be thought as the most general stance that subsumes all these: Taking a theoretical stance toward a certain question is simply taking it to be a matter to be resolved by *evidence.*

What about a deliberative stance then? We rarely come to know our own actions and beliefs through *evidence.* On the contrary, we often find evidence in the usual sense completely *irrelevant* to the question about our own actions and beliefs. For instance, a psychoanalyst comes up to O'Leary at the moment of his decision, and predicts that he will not do laundry. But that will hardly affect O'Leary's opinion of whether he will do it or not, at least at the moment of his deliberation. Our knowledge of our own actions and beliefs is supposed to

[39]Moran (1997), 154, my emphasis.

[40]Dennett (1981a) uses the notion of a stance in a similar way, for what he calls a "predictive strategy."

be *immediate* in that it is not based on evidence. But this doesn't mean that no consideration is taken to be relevant to such questions. For example, suppose that O'Leary learns that operating the washing machine at this point is likely to cause a gas leak. O'Leary will probably take this consideration (about the risk of operating the machine, not about his believing it) to be relevant to the question whether *he will do laundry.* Or suppose that Daniels realizes that the washing machine has been out of order for a while. Daniels may take the consideration (about the status of the machine) to be relevant to the question whether *he will believe that O'Leary did laundry.* In general, it seems, the considerations that the agent takes to be *reasons for* his actions and beliefs seem to be relevant considerations to resolve questions *about his own actions and beliefs.* (Or perhaps that's what makes those considerations reasons.) In the same spirit, Moran characterizes a deliberative stance as "the deferral of the theoretical question 'What do I believe?' to the deliberative question 'What am I to believe?'."[41] We can say the similar for actions: "the deferral of the theoretical question 'What will I do?' to the deliberative question 'What am I to do?'." So it seems that we can identify a certain distinctive *pattern*, loose as it may be, governing what considerations an agent takes to be relevant to the questions about his own actions and beliefs. That is, we can meaningfully speak of a distinctive stance—a deliberative stance.

The claim that this gives a *complete* story of the special way we gain *knowledge* about our own beliefs and actions must be controversial, although Moran himself is committed to this claim (at least in the case of beliefs). But what I am committed to is a weaker claim about the way a person sees *relevance* between propositions. In most cases, a person's taking one proposition to be

[41]Moran (2001), 63.

relevant to another is a matter determined by *what the person believes*, but not always. O'Leary and Daniels agree on everything about what the world is like, but assuming they are normal, they will differ in their ways of seeing relevance between propositions. Suppose that they both somehow learn that the egg Daniels is about to break into the bowl to make the omelet is rotten. Perhaps both O'Leary and Daniels will conclude that Daniels *ought not to* break the egg into the bowl. But it is only Daniels who will take that to be directly relevant to the question whether Daniels will break the egg into the bowl. O'Leary may well notice the relevance, if he knows that Daniels also learned this, and that Daniels is reliably rational. However, all those considerations will be empirical *evidence* for O'Leary, on the basis of which he infers what Daniels will do. The structure of reasoning for the two looks completely different. O'Leary takes a deliberative stance toward O'Leary's action, but Daniels takes a theoretical stance toward O'Leary's action.

Now my suggestion is this: One's taking a deliberative stance toward one's own actions and beliefs in this sense is the doxastic feature that we have been looking for, and *identifying oneself as X is (partly) a matter of taking a deliberative stance toward X's actions and beliefs* (only partly, assuming that there are other components of self-identification).

How does this help understanding Moore's paradox? To take a deliberative stance toward one's own beliefs is deferring a theoretical question "What do I believe?" to a deliberative question "What am I to believe?" And as Moran says, "in the case of the attitude of belief, answering a deliberative question is a matter of determining what is true."[42] If one fails to notice the relevance of the truth of p to the question whether he believes p, he is failing to take a

[42]Moran (2001), 63.

deliberative stance toward his own beliefs. He is treating his own beliefs just like any other person's beliefs. We can understand why one who fails to take a deliberative stance toward one's own beliefs can be said to be "alienated" or "detached" from oneself. Wittgenstein, for example, considers whether we can conceive a situation in which a person can say things like "It is raining, but I don't believe it." And he doesn't deny it is possible, but says, "we should have to fill the picture out with behaviour indicating that two people were speaking through my mouth."[43]

Now turn to Perry's problem. Recall that from the decision theoretic point of view, both *acts* and *states* are different ways of *partitioning* the space of *doxastic* possibilities. *Any* partition within one's doxastic possibilities represents the subject's *uncertainty* about what the world is like. So we can speak about different stances one takes toward such partitions. O'Leary is uncertain which state the actual world belongs to, but he will take the uncertainty to be a matter to be resolved by evidence. That is, he takes a theoretical stance toward states. O'Leary is also *uncertain* of which *act* the actual world belongs to, but the way he takes that question to be resolved is completely different: He takes a deliberative stance toward his own *acts*. Then we can define the notion of *acts* in this way: Suppose that a person takes a deliberative stance toward a certain proposition, while he doesn't take a deliberative stance toward any proposition that is properly included in it. Then we can take that proposition to be *an option* for the agent.[44]

Lastly, let us consider whether our account satisfies the intuitive assump-

[43]Wittgenstein (1974), 192

[44]Cf. Stalnaker says: "I want to suggest that the difference between active and passive knowledge is centrally involved in what it is for an agent to think of past and future actions as her/his own" (Stalnaker 1999b, 304)

tions about self-identification that we put forward earlier. First, we said that it should be something that determines a *center* in each of his doxastic possibilities. Daniels will take a deliberative stance toward beliefs and actions of *a particular person*, and doubtless that person is an inhabitant in each of his doxastic worlds. That person we can take to be a center in the doxastic world. Consider the question in a formal setting. Can we read off centers from an agent's decision matrices? Perhaps not from a single decision matrix. For, in our earlier case, O'Leary takes himself to have control not only over his body but also over the washing machine, but obviously he doesn't identify himself with O'Leary-*cum*-the-machine! But suppose that we can consider what options O'Leary would conceive himself as being confronted with in various actual or hypothetical decision situations. (For example, we consider what O'Leary would do if he learned that the washing machine was out of order.) Then, I think, we can be assured that a center can be determined.

Second, we said that self-identification should be something that one can get *right or wrong*. As far as stances concern "ways of coming to know," we can make good sense of a *success* or *failure* of a stance in a particular case. For a certain way of coming to know may fail to lead to knowledge in particular cases. If self-identification were a matter of believing an identity proposition of the form "I=TN," then the success or failure of self-identification would be a simple yes-no question, that is, truth or falsity. But when understood in our way, it can be a much more complex matter; one can fail in self-identification *in different ways*, and *to different degrees*. And I think this fits better with our intuition. For example, it can be a relatively local matter, as in believing something expressed by Moore-paradoxical sentences. Perhaps we can imagine a person who fails in self-identification more extensively; for example, by always failing to take a

deliberative stance toward his beliefs concerning a certain subject matter. Or consider a completely different kind of failures: Suppose that mad Heimson takes a deliberative stance toward *Hume's* actions and beliefs, but not toward Heimson's actions and belief. But taking a deliberative stance toward Hume will fail to get him knowledge of Hume's actions and beliefs, even if it yields some true beliefs by luck. (Or if it really gets him *knowledge* of Hume's actions and beliefs, we may have to conclude that he is right in identifying himself with Hume after all.)

1.9 Back to the Linguistic Puzzles

Both Perry's problem and Moore's paradox were originally formulated as problems concerning some puzzling linguistic phenomena. I argued that the sources of both problems lie in self-identification. I think we made some progress. But it may seem that our way of dealing with the issue makes it a little harder to account for these linguistic puzzles. So let me see what I can say about them.

One version of Perry's problem goes in this way: JP's saying "I have a class to teach," explains why he is leaving for the campus. But suppose that he instead says "JP has a class to teach," the initial explanation seems to lose the explanatory force (or at least we can easily contrive such a situation).

The case of Moore's paradox seems to raise a more urgent problem. Our account explains why it is wrong for a person to believe something expressed by Moore-paradoxical sentences; it is treating one's own beliefs just like others' beliefs. But in a way, this diagnosis is too weak. Moore-paradoxical sentences, even in such a person's mouth, sounds no less contradictory. I doubt that anything less than contradiction (semantic or pragmatic) can do full justice to this

strong intuition. Perhaps one may want to turn to a *hybrid* account: believing what's expressed by Moore-paradoxical sentences is pretty bad, but asserting it makes it *worse*, because of the nature of assertion. But I think it would be better if we could give a more unified account.

Let me close the chapter by describing two different ways to apply our findings to these linguistic problems, in a rather informal way.

(1) The first one, which I prefer, brings in some *pragmatic* considerations. Communications occur rarely in the vacuum, but against some background assumptions, or *presuppositions*. Some of those presupposition will concern some accidental features of the situation, but there may well be also presuppositions of more general kinds. That all participants of the conversation are rational might be one of them. Another general assumption that I'd like to hypothesize is involved is this: that each participant is capable of *identifying him/herself as a particular participant*.[45] To this, let's add another bit: When an ordinary person (as opposed to, say, an ATM) refers to a certain person X with "I," the person *identifies himself with* X; that is, he takes X's actions and beliefs to be something he has rational control over. If this is correct, then saying things like "It is raining but I don't believe it" will directly contradict with what's presupposed. It is a "pragmatic contradiction."

Turn to the problem of the essential indexical. JP's saying "JP has a class to teach" will lack the explanatory force of "I have a class to teach." Why not? If JP identifies himself as the person he is referring to, then, other things being equal, it would be most natural for him to say "I have a class to teach," to express that he has a class to teach. The fact that he instead uses the proper name "JP" to

[45] I believe that the latter is *entailed* by the former. That is, rationality (in its full-blown sense) entails the ability of self-identification. I will talk more about this issue in the next chapter ("Freedom and Self-Knowledge").

refer to him seems to *implicate* that JP does not identify himself as the person he is referring to after all. But this information is necessary for the explanation to be complete. For we need the information about what the alternative options the agent takes himself to be confronted with.

(2) The second one is a Fregean approach. I suggested that it was generally taken for granted by ordinary language users that "I," when used by an ordinary person, refers to the person the speaker identifies himself with. Then why not just take self-identification to be somehow constitutive of the *meaning* of "I," and some such thing to be an constituent of the thought expressed by sentences containing it? Perhaps we should enrich the traditional picture of belief with *modes of presentation* or *concepts* associated with individual beliefs (either as a part of what is believed or in some other way). Note that on anyone's account, this concept involved in "I" can't be understood as *descriptive meaning*; rather, it will have to be something like a *disposition* or *ability* to take a deliberative stance toward one's actions and beliefs, which is not communicable and private in a sense.[46]

A person who says "It is raining, but I don't believe it" lacks a complete grasp of the meaning of "I," since he fails to take a deliberative stance toward his own beliefs in this particular context. Similarly for Perry's problem. As we say that JP believes that *he* has a class to teach, we are not only attributing to him a propositional belief; we are also attributing to him certain abilities or dispositions, which involves the idea what he takes himself to have control over.

In this way, we may be able to make self-identification a matter of concepts or modes of presentation. But once we become clearer on what this consists in,

[46]I think this is essentially Evans's positive view in Evans (1982). According to Evans, "'I'-Idea" involves certain abilities or dispositions of mental or bodily self-ascriptions (i.e. self-knowledge).

I am inclined to think, there isn't much point in making this sort of move.

CHAPTER 2

On What the Two Gods Might Not Know

2.1 Introduction

The literature on *self-locating beliefs and knowledge* abounds with examples. To name only a few: There is David Kaplan's case of a man's seeing reflected image of a man whose pants are on fire, without realizing that he himself is the man. There is John Perry's case of amnesiac Rudolf Lingens who is lost in the Stanford Library and reads a lot of books, including a biography of Lingens, but still doesn't know that he himself is Lingens.[1] And then there is David Lewis's story of the two omniscient gods:

> Consider the case of the two gods. They inhabit a certain possible world, and they know exactly which world it is. Therefore they know every proposition that is true at their world. Insofar as knowledge is a propositional attitude, they are omniscient. Still I can imagine them to suffer ignorance: neither one knows which of the two he is. They are not exactly alike. One lives on top of the tallest mountain and throws down manna; the other lives on top of the coldest mountain and throws down thunderbolts. Neither one knows whether he lives

[1] Kaplan (1989), 533, and Perry (1977), 17.

> on the tallest mountain or on the coldest mountain; nor whether he throws manna or thunderbolts.[2]

Among these and other examples, I am especially interested in the last, and, of course, the light it throws on our understanding of self-locating beliefs and knowledge in general.

Why this particularly outlandish and problematic case, one might ask, not other more (literally) down-to-earth cases? Because I believe it has virtues that other examples lack. In more mundane cases of self-locating beliefs, it seems to me, two conceptually distinct problems are entangled: the problem of explaining their propositional contents, and the problem of explaining the distinctive element of self-locating beliefs not captured by them. In contrast, the case of the gods, by assuming propositional omniscience, enables us to deal with the second problem in isolation. Moreover, it has a potential to teach us something very substantial about the *nature* of the distinctive element thus isolated. The story says that the gods lack certain *knowledge*, despite their propositional omniscience. If this is right, then we may take the problem of self-locating beliefs to be *the problem of explaining this distinctive kind of knowledge*, which is apparently non-propositional.

What kind of knowledge could this be? According to some philosophers, notably Lewis, the knowledge that the gods lack is special in that it concerns a distinctive kind of information, which may be called *subjective information.* But I want to develop and defend a neglected position, which I will call *the ability hypothesis about self-locating knowledge*, obviously drawing on Lewis's ability hypothesis about *knowing what an experience is like.* According to this position, what the gods might lack despite their propositional omniscience is

[2]Lewis (1979), 139.

not any special kind of information, but a cluster of abilities, more specifically, the abilities of *self-knowledge.* And in general, such abilities are constitutive of self-locating knowledge.

I start with discussing the significance of the case of the two gods. Next I will introduce the two hypotheses that purport to explain the gods' lack of knowledge. And then I will argue why the influential accounts of self-locating knowledge held by Lewis himself, John Perry, and Robert Stalnaker are not very satisfactory. In the last part of the chapter, I will return to the ability hypothesis, and discuss some complicated problems that the case of the gods raises, and how the ability hypothesis may explain them.

2.2 The Problem of Self-Locating Belief

We begin with the so-called *traditional doctrine of belief and knowledge.* According to it, belief or knowledge is a relation between a subject and a proposition, where a proposition is understood as an abstract object that has an absolute truth condition. Following Lewis and Stalnaker, the two main figures of our discussion, I will assume that a proposition *is* a truth condition, which in turn can be identified with *a set of possible worlds.* The resulting picture is this. A state of belief or knowledge is represented as a set of possible worlds—doxastically or epistemically possible worlds. To believe a certain proposition is for it to be true in (or include) all those doxastically possible worlds. A change of belief and knowledge is understood in incremental terms; in particular, learning something is ruling out some possible worlds that were previously compatible with a state of belief or knowledge.

Self-locating beliefs are beliefs that *locate oneself* (*qua* the subject of belief)

in the world as the subject takes it to be. For example, if someone believes that he himself is David Hume, he identifies himself as Hume among numerous others in the world as he takes it to be. Or if someone believes that his own pants are on fire, he locates himself in a class of people whose pants are on fire. Everyone seems to agree that the traditional conception lacks a resource for representing self-locating beliefs properly. But it is not so easy to pin down exactly what is the problem. One popular strategy is this: We imagine someone who lacks the knowledge (or belief) of who he is, and consider what this lack of knowledge consists in. I will first consider a familiar use of this strategy, and state why I think it fails to bring out what is distinctive about self-locating knowledge, to motivate our study of the case of the gods.

Consider this example from Perry:

> An amnesiac, Rudolf Lingens, is lost in the Stanford library. He reads a number of things in the library, including a biography of himself, and a detailed account of the library in which he is lost. ... He still will not know who he is and where he is, and no matter how much knowledge he piles up, until that moment when he is ready to say, "... *I* am Rudolf Lingens."[3]

Perry and many others seem to think that Lingens's change of belief is not captured by acquiring a new propositional belief or ruling out possibilities. The point of putting Lingens in a library is to emphasize that the kind of knowledge that Lingens could acquire through reading books can't resolve his ignorance. This seems certainly right, but why should we think that only book learning is propositional? Presumably, it will take perceptual or testimonial knowledge for him to get out of his predicament. But can't these be understood in terms of ruling out possibilities?

[3]Perry (1977), 17.

The story is often accompanied by the following line of reasoning: Lingens's change of belief lies in his being in a position to say "I am Rudolf Lingens." But both "I" and "Rudolf Lingens" here are supposed to be *directly referential*.[4] So it seems to follow from this that "I am Lingens" expresses a *necessary* proposition that Lingens = Lingens, which Lingens of course knew all along. Therefore, it is concluded, Lingens's change of belief cannot consist in coming to believe a new proposition.[5] So the problem that self-locating beliefs raise is to account for this kind of change of belief, which is allegedly non-propositional. This sort of reasoning has been popular, but looks to me clearly fallacious. The fact that "I" and "Lingens" are directly referential only tells us that their referents are *all* they contribute to determining propositions expressed by statements containing them. But the way in which statements containing directly referential terms determine propositions may well be more complex than what is implied in the above reasoning.

Besides, apart from these theoretical points, I think that it is intuitively compelling that Lingens's change of belief is captured by his ruling out some of possibilities that were compatible with his belief. (And our intuition should have a real bite in this sort of debates.) Consider what the world is like according to his prior belief: There are *two* different persons: Rudolf Lingens, the biography of whom he's just read, on the one hand, and the guy who's just read the particular copy of Lingens' biography, and is standing between those particular bookshelves, wondering who he is, and so on, on the other. But his posterior belief will include only those worlds in which the two persons are one and the same. That is, some possible worlds that were previously compatible with his

[4]Kaplan (1989).
[5]E.g. Lewis (1994), 317.

belief are ruled out. And it is reasonable to identify what Lingens expresses by "I am Lingens" in this context with the set of possible worlds thus ruled out. Of course, there remains the difficult problem of explaining how statements containing "I" come to express such propositions. But this problem seems to be subsumed under a more general semantic problem of explaining how identity statements containing directly referential terms can come to express contingent propositions; for example, "Hesperus = Phosphorus," or "You are Rudolf Lingens."[6]

So I conclude that this sort of cases, or this particular way of setting up the problem, fails to bring out what is distinctive about self-locating beliefs or knowledge.

Lewis's case of the two gods is similar in structure to the case of Lingens, but it has a potential to do its job much better. The gods, one on the tallest mountain throwing down manna, and the other on the coldest mountain throwing down thunderbolts (call them Castor and Pollux, respectively), are supposed to have narrowed down their epistemic possibilities into a single one, but still don't know who they are. The only relevant difference between Lingens and the gods seems to be that the gods are more knowledgeable than Lingens. But this makes a big difference. We could interpret Lingens's change of belief in terms of ruling our possibilities, that is, propositional knowledge, only because there remained possibilities for him to rule out. But the case of the gods is different; *ex hypothesi* there is no further possibility to rule out. So their predicament must lie in something other than failing to rule out certain possibilities. In this way, the case of the gods, if really coherent, seems to *isolate* what is distinctive about self-locating belief not captured by the traditional doctrine. In fact, the

[6]What I said in the last few paragraphs is much influenced by Stalnaker (1981; 1999a; 2006).

story of the gods does more than that, since it says something very substantial about the nature of what is thus isolated: What the gods lack is something of *knowledge*. So we come to have a much refined formulation of the problem of self-locating belief and knowledge: It is to *account for this distinctive kind of knowledge* that even the omniscient gods might lack.

We gain these benefits, however, at a significant cost. We had to conceive a more extreme situation to isolate what is distinctive about self-locating belief. But by doing so, it seems that we came too far from the reach of intuition. I am inclined to distrust those who claim they simply intuit its coherence, or its incoherence for that matter. We are invited to conceive an extreme condition of omniscience, to which none of us has ever gone even close, and in trying to conceive this sort of situations, we tend to easily lapse into mistaking the real one for its lookalikes. Here are some examples: The brains of the gods are taken out of their skulls, shuffled, and then put back into the skulls arbitrary chosen, so that each god doesn't know where his brain is put in. Or the gods are wondering: "I wonder whether *this* experience [pointing to a particular token experience he is having] is occurring in Castor's brain or Pollux's brain." But all these situations are most naturally modeled by two distinct worlds. As the gods are omniscient, they know which brain is in which body, and which experience occurs in which brain. The story invites us to imagine a much stranger situation.

Those who thought about it a little more cautiously haven't reached a consensus either. For example, Lewis and Stalnaker say, respectively:

> Surely their predicament is possible. (The trouble might perhaps be that they have an equally perfect view of every part of their world, and hence cannot identify the perspectives from which they view it.)[7]

[7]Lewis (1979), 139.

> It is not obvious that the coherence of this story will survive close examination (can different agents perform different actions, without realizing, as they act, which one of them is the agent of which action?) ....[8]

It is interesting to contrast what they say in parentheses, to support their conviction and reservations, respectively. Intuitively, a primary source of self-locating knowledge seems to be *perception*, since it in its usual form is "perspectival" in some sense. But as Lewis notes, we can imagine a non-perspectival kind of perception; perhaps the gods are seeing the world only through satellite images with high resolutions. Stalnaker, on the other hand, draws our attention to another potential source of self-locating knowledge: *actions*. Notice that the gods are supposed to be agents who throw down manna and thunderbolts. But as Stalnaker points out, there is something very much odd about the idea that an agents can perform a certain act without realizing that he himself is doing the act.

So our situation is this. The case of the gods can do some things that other mundane examples can't do, but unlike the latter, its status as an *intuitive datum* seems dubious. How should we proceed then? I suggest that instead of taking it as a datum to be explained by an account, we proceed by the so-called method of *reflective equilibrium*. We may start with the *assumption* that such a god might not *know* who he is, keeping in mind that it is something eventually to be substantiated. And we set out to build a hypothesis that may explain the distinctive kind of knowledge. This hypothesis will have to be tested against intuitions about the case sharpened by this process of theorization. If anything, what should be regarded as a datum is our initial ambivalence about the case, and this also will have to be explained. If the process reaches a point of equi-

[8]Stalnaker (2008), 56. See also the footnote that follows the passage.

librium, then we may conclude that we have a plausible account of the god's predicaments, and also self-locating beliefs and knowledge in general, to the extent that the case of the gods succeeds in isolating the distinctive element of them.

So let me start by introducing two different hypotheses that purport to account for the omniscient gods' lack of the distinctive kind of *knowledge*.

2.3 Two Hypotheses: Information or Abilities?

The case of the gods is analogous to Frank Jackson's case of black-and-white Mary in certain important respects.[9] I think there is much to learn from philosophers' (in particular, Lewis's) ways of dealing with the latter. Let me review the story first.

Mary is released from the black and white room after having been confined in the room since her birth. When Mary gets out of the room, she faces colored objects for the first time. She is supposed to *learn* something, that is, come to *know what it's like to experience colors* (call such knowledge *phenomenal knowledge*). This story is originally designed as an argument against *physicalism*. It is stipulated that Mary, while in the room, learns a lot about physical sciences, including perceptual psychology, neuroscience, and psychophysics, or whatnot; we can even assume that she became omniscient as far as physical knowledge is concerned. But if she still learns something when she gets out of the room, then the learned knowledge must concern some nonphysical facts. Thus, the argument concludes, physicalism is false.

However, Lewis acutely observes that the issue of physicalism might be a

[9]Jackson (1982).

red herring. For the argument doesn't seem to make essential use of the fact that it is *physical* information. Lewis writes:

> Let *parapsychology* be the science of all the non-physical things, properties, causal processes, laws of nature, and so forth that may be required to explain the things we do. Let us suppose that we learn ever so much parapsychology. It will make no difference. Black-and-white Mary may study all the parapsychology as well as all the psychophysics of color vision, but she still won't know what it's like.... If there is such a thing as phenomenal information, it is ... independent of every sort of information that could be served up in lessons for the inexperienced.... Therefore, phenomenal information is not just parapsychological information, if such there be. It's something very much stranger.[10]

Lewis calls the hypothesis that Mary acquires a "stranger" kind of information when she gets out of the room "the hypothesis of phenomenal information." He rejects such an account, and suggests that what Mary acquires when she gets out of the room is not any sort of information, but rather certain *abilities* or *knowing-how*. "The Ability Hypothesis says that knowing what an experience is like just *is* the possession of these abilities to remember, imagine, and recognize."[11] Such abilities cannot be acquired through book learning, or any sort of lessons, but only through actually having experiences. And that's why it is only upon actually having experiences of colors that Mary acquires phenomenal knowledge.

There is a clear parallel in structure between the case of Mary and the case of the gods. Both contrive extreme situations to separate out what are distinctive about phenomenal knowledge and self-locating knowledge. And they are designed so as to make it difficult (if not impossible) to construe the lacked or acquired knowledge as a matter of propositional ignorance or learning. Some

[10]Lewis (1988), 281.
[11]Lewis (1988), 288.

philosophers appropriated this analogy in the hope that our understanding of self-locating knowledge might throw light on phenomenal knowledge.[12] But I want to reverse the strategy. I believe that the ability hypothesis about phenomenal knowledge (with some modification we will make later in this section) is plausible, or at least that it adequately explains the "cognitive" aspects of it (but what else is there to phenomenal *knowledge*?). But even if it ultimately turns out to be wrong for some reason, I think, the idea behind the ability hypothesis about phenomenal knowledge is completely general and sound, and so can be plausibly applied to self-locating knowledge.

(It is interesting that Lewis himself has never utilized this analogy. Lewis says, "if it is possible to lack knowledge and not lack any propositional knowledge, then the lacked knowledge must not be propositional."[13] He said this for the knowledge that the gods lack, but could have said the same for the knowledge Mary lacks in the room. However, his ultimate solutions diverge: As we saw, he thinks that the knowledge Mary lacks is *not information but abilities*, but as we shall see, he thinks that the knowledge the gods lack concerns *non-propositional information.* Of course, there is no incoherence here for all we know, but we may blame him for neglecting some obvious alternatives.)

Corresponding to the hypothesis of phenomenal information, we have *the hypothesis of subjective information.* By substituting "phenomenal" and "physical" with "subjective" and "objective" respectively in Lewis's formulation of the hypothesis of phenomenal information, we get a decent formulation of the latter:

> Besides [objective] information there is an irreducibly different kind of information to be had: [subjective] information. The two are inde-

[12] E.g., Stalnaker (2003a; 2008) and Perry (2003).
[13] Lewis (1979), 139.

> pendent. Two possible cases might be exactly alike [objectively], yet different [subjectively]. When we get [objective] information we narrow down the [objective] possibilities, and perhaps we narrow them down all the way to one, but we leave open a range of [subjective] possibilities.[14]

The ability hypothesis about self-locating knowledge, in contrast, says that what the gods lack is not such a "strange" kind of information, bur rather a cluster of certain abilities. And it claims that in general, self-locating knowledge is partly a matter of abilities not reducible to having information (partly, because in most ordinary cases, it also involves propositional knowledge, as in the case of Lingens). *What abilities*? Here is the basic idea, which will be fleshed out further later.

Lewis at one point describes an agent that lacks self-locating knowledge as "an agent strangely lacking in self-knowledge."[15] I think the mention of *self-knowledge* is quite apt (whatever Lewis himself meant by it). Philosophers usually mean by it knowledge of oneself gained in peculiar "first-personal" ways. For example, I can know what I believe without observing my behaviors, while others need to infer it from what I do or say. Or I can know that my legs are crossed, without looking at the position of my legs. The exact nature of such methods of self-knowledge is up for discussion, but it should be uncontroversial that we do possess certain *abilities* that we exercise only to know facts about ourselves. A crucial observation is this: *self-knowledge gained by these special ways is always self-locating knowledge.* If this is right, then bits of self-knowledge will be enough to resolve the gods' predicament, especially because they are omniscient. Suppose that Castor comes to know that his own legs are crossed, in the first-personal way. But as he is omniscient, he already knows

[14]Lewis (1988), 270.
[15]Lewis (1981), 308.

whose legs are crossed at that moment (or if not, he can easily find it out with his satellite eyes). So if it is only Castor's legs that are crossed at that moment, then his predicament is resolved, unless Pollux's legs are crossed at the same time. (But what if this proviso always fails? We will discuss such a case later.) So in order to envisage the gods' epistemic situation coherently, I think, we need to assume that *they are lacking in the abilities of self-knowledge*, or that *for some reason, they are not in a position to properly exercise those abilities.*[16]

I think all this so far should be agreed by everyone. But the hypothesis of subjective information and the ability hypothesis will differ in *what* it is that is acquired through those abilities of self-knowledge. The hypothesis of subjective information would say that it is *because* something goes wrong with their abilities of self-knowledge that the gods lack distinctive subjective information.

[16]My thought here is much inspired by Daniel Dennett's brief but insightful discussion. He considers a boat equipped with a TV set showing boats' location on a river, instead of a radar system. He says:

> What good would this do? If you were lost in the fog and looked at the television screen, you would know that one of those many moving blips on the screen was you—but which one? Here is a case in which the question "which thing in the world am I?" is neither trivial nor impossible to answer. The mystery succumbs to a simple trick: Turn your boat quickly in a tight circle; then your blip is the one that traces the little "O" on the screen—unless several boats in the fog try to perform the same test at the same time. (Dennett 1992, 427)

The predicament the boat faces with is analogous to the gods' predicament. The method Castor can use to find out which god he is is basically the same as the method the boat can use to find out which boat it is among many blips on the screen. Dennett realizes that he needs to appeal to self-knowledge:

> And how do we know that *we* are doing something? Where do we get the initial bit of self-knowledge we use for this leverage? This has seemed to be an utterly fundamental question to some philosophers ..., and has generated a literature of surpassing intricacy. If this is a substantial philosophical problem, there must be something wrong with the "trivial" answer (but I can't see what): We get our basic, original self-knowledge the same way the lobster does; we're just wired that way (Dennett 1992, 428, fn. 2).

While I am mostly sympathetic to what Dennett says here, I find his remark about the "literature of surpassing intricacy" a bit hasty (he is referring to the works by Lewis and Perry). For what these philosophers are concerned with is not *how* we know it, but *what* it is that is thus known. In a way, what I am trying to do throughout this chapter is to explain how an answer to the first question can resolve, or dissolve, the second question.

But the ability hypothesis rather says that the information they would acquire by exercising those special abilities is just the same old *objective* information about a particular god. If Castor expressed what he learned in a special way by saying, "My legs are crossed," it would have the same information Pollux's saying, pointing to him, "Your legs are crossed." There might be a sense in which objective information that Castor gained in the first-personal way is *subjective* for himself, but it differs from other information neither *in kind* nor *in its subject matter*, but only in *how* it is acquired or related to other pieces of information he possesses. (Compare: We may call some of information we possess "scientific," but presumably we don't mean something about the subject matter of the information; for what can't be a subject matter of science? It rather concerns how it is arrived at, or how it is related to other pieces of information.) The omniscient gods may be said to lack self-locating knowledge if the knowledge they have about either god fails to be subjective in that sense.

I will say more about the ability hypothesis about self-locating knowledge later. For now, let me discuss one very important concern common to Lewis's ability hypothesis about phenomenal knowledge and my ability hypothesis about self-locating knowledge. The discussion will bring out in what sense those abilities that we claim are constitutive of the kinds of knowledge in question can be called genuine *knowledge*.

Stalnaker raises the following objection to Lewis's ability hypothesis about phenomenal knowledge. (William Lycan raises a similar objection; so let's call it the Lycan-Stalnaker objection):

> But it is not clear that one can understand the relevant abilities except in terms of some notion of the intentional content of some of the things that the abilities are abilities to do. In one sense it is easy for me to imagine what it is like for the cockroach; I can, for exam-

> ple, imagine that scrambled eggs taste, to it, the way Vegemite tastes to me. If this does not count as an exercise of the ability, it must be because it doesn't get it right—because this is *not*, in fact, how scrambled eggs taste to a cockroach .... But if the ability in question is *the ability to get it right*, then it is *an ability that must be explained in terms of a kind of intentional content.*[17]

When Mary is released and sees (say) a red tomato, it seems, she will get what it's like to experience red *right.* Perhaps she could get it wrong if someone painted it green. The same point applies to the case of the gods. When Pollux resolves his predicament and comes to realize that he is Pollux, he will *get something right.* Perhaps we can conceive that he ends up with getting it wrong, by identifying himself as Castor. In short, both phenomenal knowledge and self-locating knowledge are something that can be evaluated as being *correct* or *incorrect.* I suspect that this is indeed a reason why we regard phenomenal knowledge or self-locating knowledge as genuine knowledge or cognitive achievement.

Stalnaker seems to think that from this, it follows that what Mary acquires should be understood in terms of ruling out possibilities. Lycan more bluntly says: "there is such a thing as getting 'what it is like' right, representing truly rather than falsely, from which it seems to follow that knowing 'what it's like' is knowing a truth."[18] But I disagree. I want to argue that there *are* abilities that can be evaluated as being correct or incorrect in the relevant sense, but that are not reducible to a matter of information.

Let us look more closely at one of the abilities that Lewis thinks constitute phenomenal knowledge:

> You gain an ability to recognize the same experience if it comes again. If you taste Vegemite on another day, you will probably know

[17] Stalnaker (2003b), 271, my emphases.
[18] Lycan (1995), 249

> that you have met the taste once before.... Here, the ability you gain is an ability to gain information if given other information. Nevertheless, the information gained is not phenomenal, and the ability to gain information is not the same thing as information itself.[19]

The ability to recognize experiences of a certain type is "an ability to gain information if given other information." By exercising this ability, if the situation is normal, you will get *knowledge* about experiences; for example, you will be in a position to say, "I have met the taste once before." Notice that this doesn't have to involve anything like phenomenal information; it is just knowledge that one experience is of the same type as another experience. Moreover, there is no obvious reason to think that such a disposition or ability should be reducible to a matter of some information. Now here is an important point. Although such an ability iself cannot be said to be true or false, it can be evaluated as being *reliable* or not; it is reliable if *exercising it tends to get one knowledge.* We may say that you get what Vegemite tastes like *right* if you possess the *reliable* ability to recognize the taste of Vegemite.

We can generalize the point. There are abilities we exercise to *gain knowledge.* For example, the abilities of various inferences, and the ability to form beliefs on the basis of experiences are such abilities. Surely, some of these abilities are indeed reducible to a matter of propositional knowledge. But I don't see any reason to think that all are so. Let's call an ability to gain knowledge that is not reducible to propositional knowledge a *knowledge-conferring ability.* Such an ability can be evaluated as being correct or incorrect. So I want to claim that the real lesson we should learn from the Lycan-Stalnaker objection is that the ability hypothesis about certain genuine knowledge should limit itself to knowledge-conferring abilities. Our ability hypothesis about self-locating

[19]Lewis (1988), 286-7.

knowledge obviously meets this condition, but Lewis's ability hypothesis about phenomenal knowledge doesn't.

Among the abilities that Lewis says are constitutive of phenomenal knowledge are abilities to recognize and imagine experiences.[20] Our constraint tells us to drop the second; for the ability to imagine experiences is, at least directly, not a knowledge-conferring ability. Isn't this worrisome? Such an ability intuitively looks so salient in the change of Mary's situation upon having experiences, and so the proponents of the ability hypothesis including Lewis gave a central place to it. Nonetheless, I think this is as it should be. In fact, if you take it seriously that phenomenal knowledge is *genuine knowledge*, there is something odd with the idea that the ability of imagination is *constitutive* of it. What I find a more natural thing to say is that Mary acquires the ability to imagine experiences as a *result* of her *epistemic* change. Perhaps imagining experiences can be understood as a matter of "simulating" exercises of the ability to recognize experiences.[21]

(This point has some implications for the much debated topic of what may be called "self-locating imagination"; for example, imagining myself *being* Napoleon, or imagining riding a roller coaster.[22] I think what phenomenal knowledge is to "phenomenal imagination" is what self-locating knowledge is to "self-locating imagination." And I believe that the latter can be similarly understood as a matter of "simulating" the abilities of self-knowledge. Although I will not go into this issue, it will emerge later why this makes good sense.)

---

[20] Another is the ability to remember. But I think this has to be treated specially. In a way, an ability to recognize experiences of a certain type presupposes an ability to remember experiences of that type.

[21] Cf. Thomas Nagel in his famous "bat" paper says: "To imagine something sympathetically, we put ourselves ... into a state that resembles it mentally" (Nagel 1974), 446, fn. 11.

[22] E.g. Williams (1972); Ninan (2009).

So the kind of the ability hypothesis I find most plausible is a bit different from Lewis's brand of the ability hypothesis. Lewis says that his distinction between the two hypotheses is based on the *Rylean* distinction between *knowing-that* and *knowing-how*. I say that my distinction is based on the *Ramseyan* distinction between the *static* and *dynamic* aspects of belief. The idea is that in order to describe a belief state completely, we not only need to specify what the world is like according to the subject, but also how the subject's view of the world is *disposed to evolve over time*.[23] Some such dispositions are already implicit in the ways a subject takes the world to be, but arguably not all are like that. The abilities of self-knowledge most certainly belong to the latter. And they undeniably constitute very important part of the dynamic aspect of belief; just think of how much of continual update of our belief is due to self-knowledge. What the ability hypothesis about self-locating knowledge claims is that such abilities are constitutive of self-locating knowledge.

A lot more needs to be said to make the ability hypothesis about self-locating knowledge plausible. But I guess major obstacles to seeing its plausibility are other influential accounts of self-locating beliefs and knowledge, in particular, those held by Lewis, Perry, and Stalnaker. So I will spend considerable time examining and criticizing these accounts, before returning to the ability hypothesis. Let me start with an overview of the positions, drawing a map of how they compare to each other, and also to our ability hypothesis.

[23]See e.g. Armstrong (1973), ch. 1 and Stalnaker (1984), ch. 7.

2.4 Lewis, Perry, and Stalnaker

Lewis, Perry and Stalnaker all responded to the problem of self-locating beliefs by adding some extra structure to the traditional picture of belief. A primary dividing line is this: Is the extra structure located *within* contents or outside of them? Call the former *the one component view* and the latter *the two component view*, for the reason that will become clearer shortly. Lewis's view belongs to the first, and Perry's and Stalnaker's the second. I will first compare *Lewis's* one component view and the two component view of a *generic* kind, and later explain how Perry's and Stalnaker's views fit into this generic view.

One way to contrast the two views is in terms of two different ways to utilize the intuitive apparatus of *centers* in possible worlds. It seems to be an obvious fact that an ordinary subject takes a particular person among numerous others in each of his belief worlds to be *himself* (unless the subject absurdly believes "I do not exist," but let's set this aside). That someone can be nicely represented as a center in each of the subject's doxastic worlds. There are two different ways to graft this intuitive idea onto the traditional picture of belief:

Lewis's One Component View. We may think that each belief world has a center built in it. Let a *centered possible world* be a pair $\langle c, w \rangle$, where c is a person (or some object) that exists in w. From the traditional view, replace possible worlds with centered possible worlds throughout. A belief state is represented as a set of centered worlds or a *centered proposition.* Objects of belief are centered propositions, instead of propositions. A set of centered worlds also determines a unique property, that is, the property that is had by and only by those centers, and so we may identify a centered proposition with a property. When a property F is had by all those centers, let us say that the

subject *self-ascribes F*. Self-locating beliefs are understood as self-ascriptions of properties; for example, when I say "I am hungry," I self-ascribe the property of being hungry. Ordinary propositional beliefs can also be reinterpreted as self-ascriptions of properties: To believe that p is to self-ascribe a property of inhabiting a world in which p is true. We may think that *subjective information* is distinguished from objective information in that it distinguishes between centered worlds within a world. The gods' state of knowledge is simply represented as two centered worlds in a single world—that is, the lack of subjective information.

The Two Component View. A second, perhaps less familiar, way of theorizing this idea of centers in possible worlds is this. We leave a set of possible worlds as before, and superimpose on it some extra structure determining a center in each of those possible worlds. Let f be a *function* that picks out an individual c from each belief world w. Call it a *centering function*, or *self-concept* (since a centering function is a concept in the Carnapian sense). Let us say that one *identifies oneself as* c if and only if f maps each doxastic world w to c (or c's counterpart at w, if you like). If a belief state in the traditional view is represented as B, a set of possible worlds, then in this view, a belief state is represented as $\langle B, f \rangle$. Call B *the propositional component*, and f *the self-identification component* (thus, the the label "the two component view"). Someone who says, "I am hungry," not only believes the proposition that he is hungry (the same proposition that a nearby person may express by saying "you are hungry"), but also identifies himself as that person. What about the gods' predicament? A natural thing to say is that the gods' predicament lies in some glitch in the self-identification component, but, as we shall see shortly, not all

two component theorists agree with that.[24]

In both views, a belief state can be seen as determining a set of centered possible worlds. But they are not quite equivalent. For the two component view, unlike Lewis's view, doesn't count any arbitrary set of centered possible worlds as determining a belief state. For only in Lewis's view but not in the two component view, a belief state may contain two alternatives in a single world; for a centering *function* picks out a unique c in each world.

The two views introduce different notions to represent self-locating beliefs: *self-ascription* (Lewis's one component view) and *self-identification* (the two component view). But in fact, these notions, with that of belief, seem to be *interdefinable*:

(i) S self-ascribes F if and only if
  (1) S believes that X is F, and
  (2) S identifies himself as X.
(ii) S believes p if and only if
  S self-ascribes the property of *inhabiting a world where p holds.*
(iii) S identifies himself as X if and only if
  S self-ascribes the property of *being identical with X.*

So all these notions can be regarded as legitimate by either view. But the two views take "the order of analysis" to be opposite. Lewis takes the notion of self-ascription to be prior, and can dispense with those of belief and self-identification, by (ii) and (iii), and thereby achieves greater simplicity. On the

[24]As I contrast Lewis's one component view and the generic two component view, I am in no way trying to exhaust all possible positions. In particular, I am setting aside two different sorts of the one component view. First, what may be called the ontological view, which tries to solve the problem by positing "subjective facts." Second, there is the (orthodox) Fregean position, which says a *self-concept* is a *constituent* of propositions. Frege himself apparently held such a view. According to him, "everyone is presented to himself in a special and primitive way, in which he is presented to no one else," where the way one is presented to oneself is constitutive of thoughts or contents of belief (Frege 1919/1997, 333).

other hand, the two component theorist may willingly agree that the notion of self-ascription may provide a convenient way to redescribe self-locating beliefs. But he will insist that in reality, it is something to be analyzed into belief and self-identification, as in (i).

Let's now consider how Perry's and Stalnaker's views fit into the generic two component view. First, Perry says, "having [self-locating beliefs] *could* not consist *wholly* in believing Fregean thoughts," where Fregean thoughts, for our purpose, can be identified with propositions.[25] He wants to hold on to the Fregean insight that "senses" or "modes of presentation" are necessary to properly characterize beliefs. But unlike Frege, Perry thinks that we should locate them *outside* contents. It's a matter of "ways of believing." When Lingens believes that he himself is in Stanford, what he believes is the plain proposition that Lingens is in Stanford, but he believes it "in the first-personal way." We may think of "the first-personal way" of believing propositions as determining a centering function.

Stalnaker, in his recent book, proposes a "modified centered worlds account."[26] Like Lewis, he thinks that a belief state needs to be represented as a set of centered worlds. But unlike Lewis, he imposes an additional constraint upon the structure, to the effect that a belief state cannot contain two centered worlds in a single world. Because of this constraint, Stalnaker's representation of a belief state determines a *centering function.* Moreover, he explicitly says that centers are not involved in contents of belief. Rather, he says, "the role of the centers is ... to represent where, in those worlds, he takes himself to be."[27]

The agreement between Perry and Stalnaker doesn't go much deeper, how-

[25]Perry (1977), 16.
[26]Stalnaker (2008), Ch. 3.
[27]Stalnaker (2008), 54

ever. For they significantly differ in how to divide the "shares" of the two components (partly due to their different views on belief, but obviously also partly due to their different views on self-identification). For example, Perry seems to think that when Lingens comes to know that he is Lingens, he doesn't learn a new proposition but comes to apprehend what he has believed in a new way. It seems to me that this is giving too little a share to the propositional component, and too much to the self-identification component. On the other hand, Stalnaker thinks that "ordinary belief about where you are in the world is always also belief about what possible world you are in,"[28] and that in particular, what Lingens learns is a new proposition. But he goes a step further, and apply the same to the "extraordinary" case of the gods, thereby committing himself to another form of *the hypothesis of subjective information.* It seems to me that this is giving too much a share to the propositional component, and too little to the self-identification component.

The ability hypothesis about self-locating knowledge I develop sides with the two component views in locating the extra structure outside contents. But it complains that the other two component views left its nature at best obscure (Perry), or unexplained (Stalnaker). So it puts forward a bold hypothesis about the *nature* of self-identification. It claims that self-identification is constituted by the abilities of self-knowledge, and that we can see the possession of those abilities as a genuine cognitive achievement, without granting anything like subjective information. The ability hypothesis agrees with the other two component theorists that the traditional picture of belief falls short of representing a belief state completely; however, it also insists that no modification of any exotic kind is necessary. For, as I said, it says that what's left out in the

[28]Stalnaker (2008), 51.

"traditional" picture of belief is the *dynamic* aspect of a belief state, which also "traditionally" has been recognized as a constituent of a belief state. The idea is this: You describe what a person believes at a moment (that is, a set of worlds), and how it is disposed to evolve over time. Then we can read off a centering function from it.

In next two sections, I will examine Lewis's and Stalnaker's accounts in some more detail (but we leave behind Perry's view for the reason of space).

2.5 Against Lewis's One Component View

Lewis's account has been very popular, to the point that it almost deserves to be called a new received view. So for our purpose, it is very important to see what is wrong with this influential view. Unfortunately, it hasn't been very clear to both proponents and opponents of Lewis's one component view precisely *what is at stake* in the debate. I will first present some considerations that seem to be in favor of (or at least not against) Lewis's one component view. And then I will claim that a decisive consideration should be a *functionalist* one, and argue that by that standard, the two component view is a winner.

As we saw in the previous section, there is a sense in which Lewis's one component view is simpler than the two component views. Lewis would say about the latter (as he does with respect to Perry's account): "I am sure it works as well as mine, but it is more complicated. I doubt that the extra complexity buys anything."[29] Then what can the two component theorist say in response? (From this point, things I will say in the mouth of the two component theorist may not be agreed by Perry and Stalnaker.) He is inclined to share the sentiment

[29]Lewis (1979), 151.

Gareth Evans expresses in the following passage:

> 'I'-thoughts give rise to the most challenging philosophical questions, which have exercised the most considerable philosophers, including Descartes, Kant, and Wittgenstein ....[30]

The two component theorist thinks that although Lewis's view undeniably provides a nice and elegant way of describing a belief state, it is far from giving an *account* of "the most challenging philosophical problems." We need to ask what it *is* to self-ascribe properties, and an obvious first step to answering it is to analyze it into belief and self-identification. But (asks the two component theorist) what can Lewis say about the nature of self-ascriptions of properties? He just leaves it as a *primitive* notion, which can be only partially defined.[31] Do we really have an *account* of self-locating beliefs here?

But this is a bit hasty. Consider the following methodological point Lewis made in a different context:

> Not every *account* is an *analysis*! A system that takes certain Moorean facts as primitive, as unanalyzed, cannot be accused of failing to make a place for them. It neither shirks the compulsory question nor answers it by denial. It does give an account.[32]

I agree with the basic point Lewis makes in this passage (although I am inclined to think that the line between an analysis and an account is much less clear

[30]Evans (1982), 205.

[31]For Lewis, self-ascription is a species of *de re* beliefs, which in turn is explicated by self-ascription of properties. More precisely:

> *S* ascribes property *F* to *x* under *acquaintance* relation *R* iff
>
> (1) *S self-ascribes* being *R*-related uniquely to something that is *F*, and
>
> (2) (as a matter of fact) *S* is *R*-related to that thing.

Note that this definition involves self-ascription. Self-ascriptions are distinguished from other ascriptions of properties by involving a distinctive acquaintance relation: identity. Note that if what we want is an analysis of self-ascription, what we get is obviously circular. See Lewis (1979), Secs. XIII and XIV.

[32]Lewis (1983b), 20-1.

than Lewis seems to think it is). Every theory is bound to have some primitive notions, and what notion to take to be primitive in a theory can be a somewhat arbitrary matter. It would be nice for an account to take "Moorean facts" to be primitive. But then facts about self-ascriptions seem to pass that; all parties would agree that it is a quite intuitive notion. I think Lewis is also right that leaving it as primitive is not necessarily shirking giving an account of it. We account for something not only by reducing it to some other things, but also by reducing other things to it. And Lewis in effect is claiming that the notion of self-ascription precisely does that; for he says that beliefs in general can be subsumed under self-ascriptions.

The friend of Lewis's account may want to turn the table on the two component theorist, charging him of a potential fallacy. It may be common ground that self-ascribing a property entails believing a proposition. The friend of Lewis's account might think that from this, the two component theorist hastily proceeds to conclude that self-ascriptions can be *analyzable* into believing and self-identification. But in general, such a move from entailment to analysis is illegitimate. Take an analogy: It is uncontroversial that knowing, in anyone's account, entails truly and justifiably believing, but not *vice versa* (because of the familiar point due to Edmund Gettier). However, (as Timothy Williamson forcefully argues) this doesn't imply that belief is "conceptually prior" to knowledge, or that the latter is analyzable into justified true beliefs and some other condition.[33] It may well be that it is impossible that the extra condition is not explicable in a non-circular and independent way. In fact, we may think that repeated failures confirm that such an analysis is not forthcoming. Likewise, the friend of Lewis's account might press on: "Do you have anything informa-

[33]Williamson (2002). 2-5. Sec. 1.3.

tive to say about what you call self-identification, except that it is 'whatever transforms beliefs into self-ascriptions'? You said that it was one of the most challenging philosophical questions, which had exercised the most considerable philosophers. But it is not uncommon that some vexing philosophical problems turn out to be pseudo-problems only to be dissolved." Again, I admit that there is a good general philosophical point in this. (But we should also note that if we succeed in giving a substantial account of self-identification, its force will be significantly weakened.)

However, I think there are considerations that could potentially override all these. We should not lose sight of what's the point of this whole business of ascribing attitudes to subjects: It is to *explain our rational activities.* Theories of mind recognize mental events, states, faculties and so on, as they are necessary for explaining rational activities, and distinguish them *by their functions.* Let me illustrate with the help of an analogy how this simple point can potentially be used to settle the debate between Lewis's one component view and the two component view. Suppose that a (imaginary) philosopher proposes the following:

> The classical view describes mental states as relevant to rationality in terms of belief and desire. But I find this unnecessarily cumbersome. Instead, I propose the following: Let a *valued proposition* be a function g that maps a possible world w to a pair $\langle v, u \rangle$, where v takes the values 1 or 0 (meaning truth or falsity), and u measures the extent to which the person wants w to be the actual world. Then a person's mental state can be completely represented as a valued proposition. "I am sure the classical view works as well as mine, but it is more complicated. I doubt that the extra complexity buys anything."

I am confident that all would agree that this amounts to artificially conflating two separate attitudes, only to get fake elegance. Why? I think one important

reason is that the roles belief and value (or desire) play in various rational activities are so clearly distinguishable. For example, in rational actions, desire is supposed to divide possibilities into desirable ones and non-desirable ones, while belief is supposed to distinguish possible consequences of the agent's choice of action. Now suppose that belief and self-identification are functionally distinct in a similar manner. Then are we not entitled to say that the extra complexity of the two component view really does buy *something*? Perhaps Lewis's one component view, just like the above suggestion, conflates two functionally distinguishable mental features (belief and self-identification) into one (self-ascriptions), and a *centered proposition*, just like a "valued proposition," is a mere artifact of such a conflation.

We can view some of the influential arguments against Lewis's one component view (or for the two component view) in this functionalist light. Stalnaker objects that Lewis's account makes it hard to account for communication, which must be an important kind of rational activities. A straightforward explanation of it goes like this: By saying "I am hungry," I express what I believe, and you come to believe what I said.[34] So communication is simply understood as exchange of information. But how can Lewis's view account for this? The friends of Lewis's account may respond that they can help themselves to (uncentered) propositions as well in order to explain communications, since one can easily abstract propositions from centered propositions. But this, it seems to me, is to miss the real point of the objection. If you agree that we need to abstract propositions from centered propositions in certain contexts of understanding rational activities, then, according to our standard, you are admitting that there is a reason to keep belief as a separate mental state in the first place.

[34]Stalnaker (1981), 146-7 and Stalnaker (2008), 50-2.

So I think Stalnaker's objection, when understood properly, works as a powerful argument against Lewis's one component view. But I still find it little short of being sufficiently general to convince the friends of Lewis's account. First, because it addresses only a limited (albeit important) range of rational activities. Second, because it is silent about what role self-identification is supposed to play.

It is Perry's unique contribution to the subject to draw our attention to *the role of self-identification* in explaining rational actions even of a very basic pattern, which perhaps underlies virtually all rational actions. He asks, "Why should we care how someone apprehends a thought, so long as he does?" He gives the following as "a barest suggestion of an answer":

> We use senses to individuate psychological states, in *explaining and predicting action.* It is the sense entertained and not the thought apprehended that is tied to human action. ... When you and I both apprehend the thought that I am about to be attacked by a bear, we behave differently. I roll up in a ball, you run to get help. Same thought apprehended, different sense entertained, different behavior.[35]

I find Perry's insight very important, although it is not very clear how Perry's positive view is supposed to work. The two persons that Perry refers to by "I" and "you" in the story (call them John and David respectively), we may suppose, share relevant beliefs and desires. Despite that, they are motivated to act differently. Why? It is intuitively plausible to think that self-identification is doing some work here. That is, John and David act differently despite sharing belief and desire, because they identify themselves as different persons; one as the one under attack, and the other as the one in a position to run to get

[35]Perry (1977), 19, my emphasis. See also Perry (1979; 2006) for a more sustained discussion of the issue.

help. So if we can get clear on the structure of action explanation, and how self-identification plays such a role in it, it will emerge whether self-identification and belief are "functionally continuous" or not.

I took up this task in the last chapter (Chapter 1, "Perry's Problem and Moore's Paradox"). Considering the issue in Bayesian decision theory, I argued there that the two indeed play completely different roles. Very roughly, the idea is this: An explanation of a rational action presupposes that an agent takes himself to be confronted with a range of options, which he has power to actualize at will. In decision theory, this is represented as a way of partitioning the space of the subject's doxastic possibilities. I claimed that while the role of belief is to determine this space of possibilities wherein deliberations occur, the role of self-identification is to determine the partition of the agent's options. For example, John's options partition the space into the ones where *John* rolls up in a ball and the ones where he doesn't. It is because he identifies himself as John. If he identified himself as David instead, his options would partition the space in a different way.

So I conclude that these functionalist considerations provide very good reason to prefer the two component view to Lewis's one component view. As we shall see later, the fact that self-identification has a distinctive role to play in rational actions will also have some implications for understanding the gods' predicament.

2.6 Against Stalnaker's Haecceitism

According to the two component view, a belief state is composed of a propositional component and a self-identification component. The omniscient gods are

supposed to be perfect with respect to the first, and so it seems to me more than natural to think that the second must be at fault. But Stalnaker takes a somewhat "unnatural" position: Although he recognizes the need for the distinctive self-identification component, he tries to account for the gods' ignorance by a form of *the hypothesis of subjective information.* I will first say why his specific proposal is unacceptable, *according to his own lights*, and then discuss what consideration moves him to take such an unnatural position, to rebut it.

Stalnaker says:

> The case of the two gods, as I would describe it, is ... a case of ignorance of which of two indiscernible possible worlds is actual. One of these possible worlds is the actual world ..., while the other is like it except that the god who is *in fact* on the tallest mountain is instead on the coldest mountain, with all the properties that the god on the coldest mountain in fact has.[36]

This amounts to *haecceitism*, the doctrine that says that there are two distinct possible worlds that are indiscernible in every qualitative respect. There are two worlds, w and v, which are alike in qualitative respects, but Castor is supposed to be on the tallest mountain in w, while being on the coldest mountain in v, and his ignorance lies in his failing to rule out v among these. This view implies that the property *being Castor* isn't determined by his qualitative features. Let us call this property Castor's *haecceity.*

At this point, let us consider another possible form of the hypothesis of subjective information. Recall the analogy with phenomenal knowledge. A popular response to the problem of Mary is to posit some "phenomenal facts" that are not reducible to "physical facts"; there are irreducible properties of experience, or *qualia*, the information about which Mary acquires upon release. Similarly,

[36]Stalnaker (1981), 144.

we can think of a solution positing "subjective facts" not reducible to "objective facts" (although such a view is apparently much less popular than positing phenomenal facts). Stalnaker considers such a view, under the rubric of *the ontological view*—only to mock it:

> There is a property that is in fact unique to TN that he calls 'being me' and that is distinct from the property 'being TN.' Or perhaps there is an entity—his objective self—that TN calls 'me,' and that is distinct from TN. It is an objective fact, by which we mean here a fact that must be included in a complete conception of a centerless possible world, that TN has this property, but this is a contingent fact. This very property might have been possessed by SK instead of TN (or, if we put the view in terms of objects rather than properties, it is a contingent fact that TN's objective self resides in him, rather than in SK). There is a possible world exactly like the actual world, except for the fact that the self properties of TN and SK are interchanged.[37]

We can adapt this into an account of the gods' ignorance simply by replacing TN and SK with Castor and Pollux. What the gods don't know is whether their self-properties are instantiated by the god on the tallest mountain, or the god on the coldest mountain.

Now my worry about Stalnaker's haecceitism is this: Is the haecceitist move, which Stalnaker explicitly endorses, is really distinguishable from the ontological view, which he explicitly rejects? We may take the following two theses to define a self-property according to the ontological view.

(a) There is Castor's *self-property* that is in fact unique to Castor, and that he calls "being me," which is distinct from the property of *being Castor.*

(b) Castor's *self-property* might have been possessed by Pollux instead of Castor. That is, there is a possible world exactly like the actual world except

[37]Stalnaker (2003a), 259.

> for the fact that Castor's *self-property* is had by Pollux (that is, the one who possesses *being Pollux* in that world.)

But here by replacing "Castor's self-property" and "being Castor" with "Castor's haecceity" and "Castor's qualitative properties," respectively, in (a) and (b), don't we get precisely what would characterize Stalnaker's haecceitism? Compare Castor's *self-property* and Castor's *haecceity*. They both are supposed to be properties determining Castor's identity, but not determined by qualitative properties of him, and their bearer is supposed to be the referent of "I" in Castor's mouth. But then aren't haecceitism and the ontological view dreaming of the same kind of properties, only verbally different? Stalnaker says at one point, "Few are tempted to try to explain this distinctive kind of knowledge by refining our metaphysical conception of the objective world—by objectifying the self."[38] But it seems to me that by being committed to haecceitism, he sides with the few so tempted, whether he wants it or not.

(Some may think that haecceities have other theoretical roles to play, apart from explaining this kind of ignorance and other related phenomena (for example, to make sense of identity across possible worlds). That might provide a reason to doubt the equivalence of the two views, as it might give haecceities further roles that are not satisfied by self-properties. But all I need to say is that the two views are relevantly similar and subject to exactly the same objections.)

What's so bad about the ontological view then? The basic problem is that it by itself doesn't seem to solve the problem it is posited for. For suppose that Castor comes to know that Castor's self-property is had by the god on the tallest mountain. How does that help to resolve his uncertainty, unless he

[38]Stalnaker (2008), 36.

already know that *he himself* is the one who has Castor's self-property?[39] The haecceitist move faces exactly the same challenge. We can see Lewis's famous objection to the haecceitist move in this light. He questioned whether those qualitatively indiscernible alternatives can really capture the gods' ignorance. For even after Castor narrows down his epistemic alternative into a single world (among two qualitatively indiscernible worlds), the question where Castor is in that world still seemed to remain.[40]

Stalnaker once rebuffed this as question begging, but later admits that the haecceitist move needs to be amended:

> Lewis is right, I think, that the haecceitist move does not eliminate the need to link the believer to the worlds compatible with his or her beliefs, and so does not, by itself, provide an account of the states of ignorance of the two gods.[41]

The "link" here is what we call a *centering function.* His idea seems to be this: Castor's epistemic possibilities are w and v, qualitatively indiscernible worlds. Castor's centering function f_c maps w to the god on the tallest mountain, and v to the god on the coldest mountain. Pollux is in the same epistemic situation as far as his epistemic possibilities are concerned. But his centering function is a different one, f_p, which maps w to the one on the coldest mountain, and v to the one on the tallest mountain.

But, it seems to me, what's really doing the explanatory work here is not two qualitatively indiscernible worlds, but a centering function. Then can't we just have the latter do *all* the work and dispense with mysterious haecceitism

[39]Cf. Nagel (1986), 56. There is a way out for the proponents of the ontological view. They can say that only Castor can be "acquainted" with Castor's self-property. But as Stalnaker points out, this looks to be an attempt to patch a mystery with another mystery. See Stalnaker (2003b), 259-60.

[40]Lewis (1979), 140-1 and Lewis (1983a), 394, fn. 16.

[41]Stalnaker (2008), 57.

altogether? An alternative picture I have in mind is this: Castor narrowed down his epistemic possibilities into w. But he is *ignorant of which of two centering functions f_c and f_p is a correct one.* Then we don't need the two qualitative indiscernible worlds, w and v, any more. Why doesn't Stalnaker take such a position seriously, if he recognizes the need for a centering function? I think his reason is this. As we saw when we discussed the Lycan-Stalnaker objection to the ability hypothesis about phenomenal knowledge, he seems to think that *getting something right* must be explained in terms of eliminating possibilities. So if Castor is *ignorant* of which centering function is a *correct* one, that is something to be explained by his failure to rule out certain possibilities. And in order to explain that, he would say, we will have to bring back in two qualitatively indiscernible worlds.

But I claim that our ability hypothesis about self-locating knowledge provides a way to break out the circle. It says that there *are* some cognitive abilities that can be evaluated as being correct or incorrect, but that don't have to be explained in terms of ruling out possibilities; that is, knowledge-conferring abilities. So we finally return to our ability hypothesis about self-locating knowledge.

2.7 The Ability Hypothesis

We have special abilities we exercise to gain knowledge about some facts about ourselves, which in ordinary circumstances we never use to know corresponding facts about others. For example, I know that I believe or desire something, that I am having particular experiences, or thinking particular thoughts, or that my legs are crossed, in peculiar ways. I am about to claim that *all* these abilities

can be the source of self-locating beliefs. But the following two abilities of self-knowledge seem to me particularly salient in self-identification.

First, I can know *where I am located relative to some items in the world* in a peculiar way. For example, long before Lingens comes to know that he is in Stanford, he must have realized that he is in front of those particular bookshelves. How did he? Simply *by looking around.*[42] Perceiving those bookshelves seems to be enough to get him the knowledge that he is in front of those bookshelves. Another person nearby may come to know the same, that is, that he (Lingens) is in front of those particular bookshelves, but the way she knows it will be completely different; she might have observed *him* facing with those bookshelves. (Some might think that all Lingens can know by exercising the special ability is that *he is seeing* those bookshelves. And from this and his background belief that he can see only what he is facing, he infers that he is facing with those bookshelves. For our purpose, the details don't matter.)

Second, I can know *what I will do in the near future* in a peculiar manner. For example, Lingens comes to know that he will soon go downstairs to find his way out. How did he? Simply *by making decision.*[43] He might have reasoned as follows: "The circulation desk is downstairs, and a circulation desk in a library is usually located nearby its main entrance. So I will go downstairs." Another person may *predict* the same, that is, that he (Lingens) is soon going downstairs. But her reasoning will take a different form, say: "Lingens *believes* that the the entrance of the library is downstairs, and eagers to get out of the library. He

[42]Evans (1982), 231-3.

[43]Anscombe (1957); Moran (2001). See also Stalnaker (1999b). Recall we based one of our objections to Lewis's account on the idea that self-identification has a distinctive role to play in rational actions; it distinguishes doxastic possibilities in terms of what the agent has control over. This in fact imposes a constraint on an account of self-identification; it must be something that can play this role. I argued in Chapter 1 that taking oneself to have control over X's action is a matter of being disposed to know X's action by making decision.

has always acted rationally in this sort of situation. So he will go downstairs." (Again, some might think that all Lingens can know by exercising the special ability is that *he intends to go downstairs.* And from this and his background belief that there is nothing that keeps him from realizing this intention, he infers that he will go downstairs. Again, for our purpose, the details don't matter.)

Let's call knowledge and beliefs that are disposed to be gained by exercising those special abilities *self-knowledge* and *self-belief.* It is important to note that our definition of these notions does *not* concern the subject matter of beliefs and knowledge, but only the ways they are disposed to be arrived at. In particular, self-knowledge and self-beliefs don't have to be *about oneself.* For example, suppose that Heimson, being mad, always forms beliefs about Hume in the ways ordinary people form beliefs about themselves; he finds out where Hume is simply by looking around, and what Hume will do simply by making decision. These beliefs of Heimson's also count as self-beliefs under my usage of the notion (but they will be very unlikely, as a matter of fact, to be self-*knowledge*). Let us say that in this case, Heimson has self-beliefs *about* Hume. And when one is disposed to form self-beliefs about person X, let's say that the subject possesses *the abilities of self-knowledge about* X. Now my ability hypothesis claims:

> Necessarily, S identifies oneself as X if and only if S has the abilities of self-knowledge about X.

In the ordinary situations, this will get right results. Moreover, the hypothesis implies that our Heimson, who has the ability of self-knowledge about Hume, identifies himself as Hume. Notice that the right side of this biconditional can be evaluated as being right or wrong, as it can be reliable or not. The left

side of it will inherit the evaluation. Suppose that as is natural, Heimson's self-beliefs about Hume are mostly false. Then we may say that Heimson gets self-identification wrong.

The two component theorists think that self-locating beliefs are constituted by propositional belief and self-identification. So there must be intimate connection between self-knowledge and self-locating knowledge. First, our hypothesis implies that *self-knowledge is always self-locating knowledge.* This seems intuitively right. Obviously, the converse doesn't hold; that is, not all self-locating knowledge is self-knowledge. In order for Lingens to know that he is in Stanford, not in Harvard, he needs more than self-knowledge. However, the following weaker thesis seems to me plausible enough: *All self-locating knowledge is ultimately based on some self-knowledge.* Suppose that Lingens comes to know that he is in Stanford. Ask him this question: "How do you know that it is *you yourself* that is in Stanford?" It is only when Lingens reaches the point where he can appeal to self-knowledge when a further question can't be asked. For example, Lingens might say: "I know only the Stanford library has a copy of the rare biography of Rudolf Lingens, and I am now right in front of the book."[44]

[44]I think that the notorious phenomena often called *immunity to error through misidentification* (first noticed by Ludwig Wittgenstein and so named by Sydney Shoemaker (1968)) can be interpreted as providing direct evidence for our hypothesis. Wittgenstein says:

> It is possible that, say in an accident, I should feel a pain in my arm, see a broken arm at my side, and think it is mine, when really it is my neighbour's. ... On the other hand, there is no question of recognizing a person when I say I have toothache. To ask "are you sure that it's *you* who have pains?" would be nonsensical. (Wittgenstein 1958, 67)

First, I think (and Wittgenstein wouldn't deny) that Wittgenstein's point applies to all self-beliefs or self-knowledge. Second, what is said to be nonsensical is a question about self-identification; that is, "are you sure that it's *you yourself* who is *F*?" As I take it, this shows that there is a conceptual connection between self-knowledge and self-identification.

2.8 What the Gods Might Not Know

I think that, at least initially, we can distinguish *two different ways* in which the omniscient gods might lack self-locating knowledge. To explain what I have in mind, let me go back to the analogy with the story of black-and-white Mary once more. Consider the following variation on the story, due essentially to Martine Nida-Rümeline.[45] Suppose that before being completely released, Mary is first transferred to another room, whose wall is painted red uniformly. Call it the Nida-Rümeline room. It seems that she already learns something in this room, that is, what it is like to experience red, although not under that description. In the next step, when she is completely released, and sees some recognizable red objects, she learns that the experience she had earlier is an experience of red.

I think the (modified) ability hypothesis about phenomenal knowledge can explain these two steps of Mary's epistemic progress, perhaps better than any other alternatives. In the Nida-Rümeline room, Mary acquires the abilities to recognize experiences of a certain type. So she is in a position to say, "*This* experience [referring to the experience she is currently having] is of the same type as *that* experience [referring to the experience she had a moment ago]." It is only in the next step, that is, when she is released, that she can give independent descriptions to the type of experiences she is already able to recognize. Now I want to adopt the following locution as a convenient way of saying the same thing: in the Nida-Rümeline room, Mary acquires a *phenomenal concept* of red experiences, and when released, she learns what the phenomenal concept is *true of*. In fact, this is not an arbitrary way of speaking at all. In general, a con-

[45] Nida-Rümelin (1995). See also Perry (2003) and Stalnaker (2008). Lewis also observes the similar point: "One might even know what some experience is like, but not under any description whatever. ... That is what would happen if you slipped a dab of Vegemite into my food without telling me what it was" (Lewis 1988, 287).

cept is something that picks out individuals in possible worlds, by being *true of* those individuals in them. As the ability to recognize is a *knowledge-conferring ability*, it can be "correct of" a particular type of experiences. And so we may justly say that such an ability constitutes a concept.

I think we can make an analogous distinction between two different ways in which the gods might not know who they are. Again, I will help myself to the notion of a *self-concept*: the abilities of self-knowledge can pick out individuals in possible worlds by being "correct of" those individuals, and so constitute a concept. The first corresponds to Mary's situation in the black-and-white room: *The gods possess no self-concept.* This means that the gods lack the abilities of self-knowledge altogether. The second is comparable to Mary's situation in the the Nida-Rümeline room: *The gods possess self-concepts, but don't know which god their self-concepts are true of.* This means that the gods are able to gain self-knowledge about someone, but can't give independent descriptions of the one about whom he gains such knowledge. Let's examine these two ways closely in turn, in the reverse order.

First, could each god possess a self-concept, but not know which god it is true of? Before getting to the case of the gods, consider a variation on the story of Lingens. Suppose that it is because Lingens got hit by a tome that fell off from a library shelf that he lost his memory. He has been in a coma between shelves unattended for sometime, and now starts to regain his consciousness, but is still paralyzed; he can't see or hear anything, and can't move his body. He finds himself in utter darkness. We may suppose that his amnesia is of a strange kind; he retains all his propositional beliefs, but lost track of how his belief system has *evolved.* Before he lost his memory, he knew that he (that is, Lingens) is in a library. Now he retains the same propositional belief, that is, that Lingens is in a

library, but doesn't remember how he came to know it. According to the ability hypothesis, this means that he doesn't remember *himself* as Lingens. Now just as Mary in the Nida-Rümeline room has the abilities to recognize experiences, without knowing what type of experiences they are true of, I think that Lingens in this situation may retain his abilities of self-knowledge, without knowing whom these abilities are true of.

What corresponds to Mary's final release is Lingens's recovering his perceptual and behavioral abilities. Mary, when released from the Nida-Rümeline room, comes to know what her phenomenal concepts are true of. This would be basically through observing some objects the colors of which she already knows. Similarly, when Lingens recovers his perceptual and behavioral abilities, he will be in a position to exercise his abilities to self-knowledge, and so learn things about the person whom his self-concept is true of.[46] As he perceives particular bookshelves, he comes to know that his self-concept is true of the person facing those particular bookshelves, and as he decides to go downstairs, he comes to know that his self-concept is true of the person who is going to start moving downstairs at the moment. (The original story of Rudolf Lingens may be inserted at as late as this stage. Lingens will be able to accumulate a considerable amount of self-locating knowledge in this way, but still remain ignorant of whether he is in Stanford or in Harvard. That's the matter to be resolved by propositional knowledge.)

Then could the gods' situation be thought to be similar to that of Lingens on the brink of waking up from a coma? We are free to imagine so. But if that's what we are conceiving, I think, then we are departing a little from the original

[46]Strictly speaking, Lingens could have exercised *some* of his abilities of self-knowledge even in darkness. For example, he will be in a position to say, "I am the one who is thinking *this* thought."

scenario described by Lewis. According to it, they are said to be normal in other respects: in particular, they are *perceivers* who look down the earth, and *agents* who throw down manna and thunderbolts. And, as they possess the abilities of self-knowledge or self-concepts, they can easily find out who they are, by a simple trick. Castor may try to raise his right hand to see whose hand rises. If it is Castor's hand that rises, then he comes to know his self-concept is true of Castor.[47] Then wouldn't it be impossible that the gods, who are omniscient and agents and also have self-concepts, remain ignorant of who they are?

Well, I think it *is* possible, although it takes a greater contrivance to conceive such a possibility. Notice that the simple trick can resolve Castor's uncertainty only under the assumption that *only one* of the gods raises his right hand at that moment. But what if that assumption fails, not only for this particular action, but for all other actions? Again back to Mary. If the situation is normal, after Mary is released from the Nida-Rümeline room, she will be able to easily find out what some of her phenomenal concepts are true of. But suppose that a mischief were determined to *deceive* Mary, and painted all objects along her path with a single color. Then she would be unable to tell what colors her phenomenal concepts are true of. Now it would be much more difficult to deceive the gods, because they are omniscient (if someone were to deceive them, then they would already know that too). But what if *Mother Nature* played the role of a deceiver? Suppose that their world is such that Castor and Pollux are independent subjects just as any two of us are, but that strangely, their behaviors are always coordinated (as if a higher god had pre-adjusted their behaviors in advance). Castor decides to raise his right hand, but what soon happens is that both gods' right hands go up, even though they are not causally connected. And

[47]Recall Dennett's story of the boat.

likewise for every action. If that were the case, the gods would be destined to be ignorant of who they are forever.

The last scenario, despite its artificiality, has some theoretical significance. First of all, the fact that the gods needed to exercise their perceptual and behavioral abilities to know their identity tells us that some *empirical elements* may be involved in self-identification. This necessarily brings with it an element of "luck." What the above artificial scenario purports to show is how things independent of the gods can leave them out of luck. I am inclined to think that such an element of luck is inherent to the concept of knowledge. For knowledge always requires more than efforts on the knower's part, even apart from truth of what's known, and so if circumstances go sufficiently awry, a fine method of gaining knowledge may fail to get one knowledge. I think this point reinforces our claim that self-identification is a kind of knowledge.

Let's turn to the second contemplated way of the gods' ignorance; that is, *lacking self-concepts* altogether (corresponding to Mary in the black-and-white room, where she lacks phenomenal concepts). This means that the gods simply lack the abilities of self-knowledge about *anyone*. For example, he is not in a position to say, "I'm thinking this particular thought," or "I am going to throw down manna." Is this a possible situation? I am not sure. It is because of the point we made about *the role of self-identification in rational actions*. We earlier said that self-identification plays an indispensable role in rational actions. If this is right, a subject who lacks the abilities of self-knowledge, and hence incapable of self-identification, will not be capable of rational actions at all. But can we really ascribe belief and knowledge to such a being? Perhaps there is no single correct answer to this question. Some may think that belief and knowledge can be understood only in terms of their contributions to rational actions. If so,

then we would have to conclude that this second way of the gods' ignorance is impossible. Or perhaps there is a sense in which even a non-agent, such as a tree, can be a representational system, and have beliefs and knowledge at least in some rudimentary sense. If so, then we might have to conclude that Castor and Pollux are such *deistic* gods, who represent the world, but don't intervene in the affairs of the world.

So I want to conclude that the predicament of the gods is strictly not impossible. Our account does explain why it needs some stretch of imagination. Whenever we tried to contrive situations where they lack abilities of self-knowledge, the two conditions of *omniscience* and *being a rational agent* made it difficult to do so. I think this also shows how integral self-identification is to our overall condition of rationality.

CHAPTER 3

What Is the First Person Perspective?

3.1 Introduction

It is said that we may view ourselves from two different *perspectives* or *standpoints*. When seen from the third-person (external or objective) perspective, we, our mental states, and our actions, are all just part of a natural order, perhaps made of the same kind of stuffs as all other people, and subject to the same laws of nature. But when viewed from the inside, or from the first person point of view, we, our own mental states, and our own behaviors are all specially presented to ourselves. Our experiences are presented to us as having vivid phenomenology in the way no one else's experiences are. And from the first person standpoint, we regard our actions, and beliefs and some other mental states as having "authors" or owners behind them. What is this viewing from the first person perspective, and how should we accommodate it in our world view?

Thomas Nagel is one of the philosophers who made influential contributions to our understanding this subject. In "Introduction" to *The View From Nowhere*, he says the following:

> There are things about the world and life and ourselves that cannot be adequately understood from a maximally objective standpoint, however much it may extend our understanding beyond the point from which we started. A great deal is essentially connected to a particular point of view, or type of point of view, and the attempt to give a complete account of the world in objective terms detached from these perspectives inevitably leads to false reductions or to outright denial that certain patently real phenomena exist at all.[1]

But I am not sure whether Nagel is entirely free from some elementary fallacies in this and other related passages. As he says, "appearance and perspective are ... part of the world."[2] No one should deny that. We do obviously take a first person perspective, and we, our actions, and our mental states are specially viewed from that perspective. But from the patent reality of this fact, it in no way follows that things distinctively viewed from that perspective is a distinctive reality, nor that this fact that that we are specially viewed from the first person perspective cannot be properly understood from a detached objective point of view.

What I will try to do in this chapter is to give an account of what it is for a person to take a first person perspective, and how is it different from taking a third person point of view. I will very consciously try to do this "in objective terms detached from" these internal perspectives. This is perhaps not a single subject. In particular, taking a first person perspective with respect to experiences may be something quite different from taking a first person perspective with respect to one's own actions, although it would be surprising if there were no connection at all. I will exclusively focus on the latter.

A popular view that is often implicitly taken for granted says that taking a first person perspective is simply a matter of *believing* some distinctive propo-

[1]Nagel (1986), 7.
[2]Nagel (1986), 4.

sitions. I think that this view has many problems, and that the only reason why we stick with it is the lack of a plausible alternative. So I will develop and defend an alternative account. According to the view I will develop, taking a first person perspective toward one's actions is to take a *deliberative stance* toward them, which in turn is understood as ways of knowing one's own actions. I will argue that this view explains many features of the first person perspective that the belief view doesn't explain very well.

3.2 Acts Viewed from the Frist-Person Perspective

Let me start with two passages that describe the contrast between the two perspectives. The first is from Christine Korsgaard, who attributes the view to Kant:

> [A]s rational beings we may view ourselves from two different standpoints. We may regard ourselves as objects of theoretical understanding, natural phenomena whose behavior may be causally explained and predicted like any other. Or we may regard ourselves as agents, as the thinkers of our thoughts and the originators of our actions. These two standpoints cannot be completely assimilated to each other, and the way we view ourselves when we occupy one can appear incongruous with the way we view ourselves when we occupy the other. As objects of theoretical study, we see ourselves as wholly determined by natural forces, the mere undergoers of our experiences. Yet as agents, we view ourselves as free and responsible, as the authors of our actions and the *leaders* of our lives. The incongruity need not become contradiction, so long as we keep in mind that the two views of ourselves spring from two different relations in which we stand to our actions.... These two relations to our actions are equally legitimate, inescapable, and governed by reason, but they are separate.... we must view ourselves in these ways when we occupy the standpoint of practical reason—that is, when we are deciding what to do.[3]

[3]Korsgaard (1989), 119-20.

The second is Nagel's, from the chapter titled "Freedom":

> From the inside, when we act, alternative possibilities seem to lie open before us: to turn right or left, to order this dish or that, to vote for one candidate or the other—and one of the possibilities is made actual by what we do. The same applies to our internal consideration of the actions of others. ... From an external perspective ... the agent and everything about him seems to be swallowed up by the circumstances of action; nothing of him is left to intervene in those circumstances. This happens whether or not the relation between action and its antecedent conditions is conceived as deterministic. In either case we cease to face the world and instead become parts of it; we and our lives are seen as products and manifestations of the world as a whole. Everything I do or that anyone else does is part of a larger course of events that no one "does," but that happens, with or without explanation.[4]

What Korsgaard and Nagel were trying to do, as I understand them, is to emphasize the distinctiveness and legitimacy of our internal perspectives, and construct some philosophical theories *from that perspective.* But my interest is something totally different. As I said earlier, I want to know what it *is* for us to take the first person perspective on our own actions, and what it is to regard ourselves as the originators of our own actions. Is it a species of a familiar attitude, such as beliefs or experience, which has only a distinctive content? Or is it a distinctive mental attitude?

Let me first gather a few *intuitive* characteristics of the way we view our own actions, and its relation to the way we view other people's actions, all of which I believe are explicit or implicit in the above passages.

Belief-likeness. We view ourselves as the authors of our actions. This viewing resembles belief in important respects. By believing something, we commit ourselves to a certain view of what the objective world is like. Likewise,

[4]Nagel (1986), 113-4.

when I regard myself as the source of my own actions, I seem to commit myself to a certain view of how I myself am related to certain objective states of affairs—that is, my actions. Moreover, just like beliefs, it seems that such a view may turn out to be *correct* or *incorrect*, depending on how I turned out to be actually related to my actions.

Incongruity. I regard myself as the author of my action, and you, considering the same from an external standpoint, take it to be a matter to be explained by my mental states and other natural forces. But are you and I *contradicting* each other, and is one of us right and the other wrong? Although there seems to be a sense in which the two ways of viewing one and the same action are "incongruous," or conflicts, with each other, they don't seem *contradictory.*

Subjectivity. Each of us can take the first person perspective only on his or her own actions. (In fact, it's not even clear whether it makes sense to say that one takes a first person toward other person's actions; if he views someone else's action that way, it seems, he is just taking the action to be *his own.*) We inevitably view other people's actions as natural processes. We demand explanations of other people's actions, and what we are demanding is always some kind of causal explanation. This is not to deny that we may sometimes put ourselves in someone else's shoes, and *sympathetically* take an internal point of view. But that's certainly different from really taking that point of view.

Rationality. We *must* occupy the internal point of view with respect to our impending future actions. Unless you regard yourself as the source of some

> decision, deliberation can't even start. In that sense, it is a *rational requirement* to regard yourself as the author of your action.

I think that an adequate account of the first person perspective will have to explain all these features.

A popular view, which I take to be a main rival to the account I will develop, is this:

> *The belief view*: Taking a first person perspective on one's actions is a matter of *believing* some distinctive propositions about them, something to the effect that one is the author of those actions.[5]

Viewing someone's action from the third person perspective will be understood in a parallel manner: it is a matter of believing some other propositions about those actions, for example, something to the effect that those actions are causally explained by antecedent events. That is, the difference between the two perspectives is explained as beliefs of disparate contents.

Doubtless, the belief view will explain *Belief-likeness* in a most straightforward way. One may even doubt whether there is any alternative account that can explain this feature. Perhaps this explains the popularity of the view. Then how well does it do with respect to the other characteristics? Not very well, I think. First, consider *Incongruity*. As you and I view my action from different perspectives, we are having different beliefs about the same thing. They must be either contradictory or consistent. To respect *Incongruity*, it will have to say that they are consistent. But then it's not clear what could explain the sense in which the two perspectives are incongruous or in tension.

[5]For example, Nagel explicitly says, "The sense that we are the authors of our own actions is not just a feeling but a belief," although he very soon retreats by saying, "it is no intelligible belief at all" (Nagel 1986, 114). But this view is more often just taken for granted by many philosophers, especially by those who work on the issue of freedom.

Next, consider *Subjectivity*. It says that only I can view my own actions from the first person perspective. To account for this, the belief view will have to attribute an unusual feature to the proposition that is believed when I occupy the first person point of view: it is a proposition that can be believed only by me, but not by others. But this is quite mysterious if not incoherent. It is not so clear whether the belief view can explain *Rationality* either. As usually conceived, rationality constrains what a person is to do or believe *given* certain beliefs and desires. Then how can there be a (contingent) proposition such that believing it is so much as a rational requirement *simpliciter*?

These are in no way meant to be anything close to conclusive arguments against the belief view. But I hope they do give us some justification for our pursuing an alternative account.

Before we move on, let me say a few words about why I think it is a very urgent philosophical task to reject the belief view. If what's involved in the first person perspective are beliefs with some distinctive contents, we should ask *what the world should be like* for such beliefs to be true. Studying what the world should be like for such a belief to be true amounts to *metaphysics of freedom and responsibility*, which seems to me to have reached a hopeless deadlock. But if we reject the belief view, the whole project of metaphysics of freedom may turn out misguided.

Moreover, the belief view seems to lead some dubious metaphysical view such as Nagel's that we saw in the beginning of this chapter. If the belief view is true, then it will be literally true that our objective conception of the world leaves something out; it omits knowledge of "subjective facts," which make those belief constituting the first person perspective true. Moreover, conjoined with what we called *Subjectivity*, it implies that some subjective facts are bound

to be unknowable. The case for this sort of metaphysical move will be significantly weakened if our first-personal viewing is something other than beliefs.

3.3 An Agent's Options as Doxastic Alternatives

The view I will develop says, roughly, that the difference between the two perspectives is best understood as two different ways of knowing. Rather than directly plunging into the view, however, I want to say how we may be naturally led (and how I actually was led) to such a view from some simple considerations.

Recall Nagel says: "From the inside, when we act, alternative possibilities seem to lie open before us ... and one of the possibilities is made actual by what we do." We may call "alternative possibilities that seem to lie open before the agent" *options for the agent.* The picture is this: An agent is confronted with a range of options, and he "closes" one of them in some peculiar way. And presumably, what's distinctive about an agent's first person perspective consists in the peculiar manner the agent *closes* those options. But in order to know how the agent closes them, we first need to know precisely in what sense they are *open.*

I think the following should be uncontroversial.

(Dox) In order for A to be an option for an agent, then both A and not-A should be *doxastic* alternatives for the agent. Moreover, options for an agent form a partition of the space of the agent's doxastic possibilities.

For example, if ordering salad or soup in a restaurant are options for you, then you aren't sure whether you will order salad or soup. Depending on what you believe, options for you may vary. For example, if you believe that soup is always served with bread in that restaurant, your options are between ordering

salad and ordering soup with bread. Moreover, even if your belief that soup comes with bread is mistaken, your options will still be ordering salad or ordering soup with bread.

Now comes a controversial claim.

(Phy) In order for A to be an option for an agent, the agent has to *believe* that both A and not-A are *physical* alternatives.

Notice that (Dox) follows from (Phy): If you believe that A and not-A are genuinely undetermined, then both A and not-A should be your doxastic possibilities. So those who accept (Phy) will say that (Dox) is true *because* (Phy) is true. But is (Phy) true?

I don't find (Phy) intuitively plausible at all. Consider these two cases:

(d) You are about to decide between A and not-A. But an expert, whom you perfectly trust, comes along, and informs you that you will do A. What will happen?

(p) You are about to decide between A and not-A. But an expert, whom you perfectly trust, comes along, and informs you that whether you will do A or not-A is already predetermined (but doesn't say which). What will happen?

In both cases, you won't feel as if you're deprived of control over your body and became a bystander; that is, neither of the options ceases to be an option. Instead, in case (d), you will be forced to distrust the expert's opinion. That's what both (Dox) and (Phy) successfully predict. But only (Phy) predicts that even in case (p), you will be forced to distrust the expert's opinion. Would you? I wouldn't. I am in fact a determinist, believing every single action I perform is

predetermined. But I have never felt that I was forced to give up the belief, even at the moment of decision, in the way the situation such as (d) forces me to. If the objector insists that he intuits differently, we reach a standoff.[6]

Whether we accept (Phy) or not has important implications:

If we buy (Phy), then *closing* options in a peculiar manner should amount to *actualizing* those physical possiblities in a peculiar manner. This will naturally lead to a version of the belief view. That is, what's distinctive about the first person perspective lies in the agent's *believing* that those physical possibilities are *actualized* in some special way.

But if we don't buy (Phy), things completely change. The only sense in which options are open is the doxastic sense. *Closing* options in a peculiar manner cannot be a matter of actualizing possibilities, but rather a matter of *ruling out* or *eliminating* possibilities from the space of doxastic possibilities. And eliminating or ruling out possibilities from the agent's doxastic possibilities is for him to *gain belief or knowledge.* What's distinctive about the first person perspective must lie in the peculiar way we know about our own actions. I will develop this idea from the next section.

3.4 "Stances" as Ways of Knowing

There may be different ways of coming to know the same thing. The way in which I know my actions and the way in which you know my actions are completely different. To systematically develop the idea, I choose to use the notion of *stances.* What I have in mind is very close to Daniel Dennett's notion of

[6]Daivid Velleman accepts (Dox) while denying (Phy). He tries to explain away (Phy) as an illusion.Velleman (1989) Deniel Dennett also seems to accept only (Dox). See (Dennett 1984, 113)ennett.

stances as "predictive strategies."[7] Here are some of Dennett's examples of stances or predictive strategies one may adopt to predict a person's behavior:

> [D]etermine the date and hour of the person's birth and then feed this modest datum into one or another astrological algorithm for generating predictions of the person's prospects.
>
> [D]etermine its physical constitution ... and the physical nature of the impingements upon it, and use your knowledge of the laws of physics to predict the outcome for any input.[8]

The first strategy he calls the astrological stance, and the second the physical stance. Dennett's another example, which he is primarily concerned with, is the intentional stance, by which he means, roughly, a strategy predicting one's behaviors from beliefs and desires by rationalization. But once we get the idea, we can think of a host of other examples; for example, we may adopt the sociological stance, or the psychoanalytic stance, to predict a person's behavior.

So what exactly are predictive strategies? Dennett just gives examples in the form of "instructions," as in the above passages. But I want to characterize this notion of stances in a little more general and abstract way than Dennett does. Let's take the following as a basic notion: A subject *takes p to be relevant to q*, where both p and q are propositions. What I mean is this:

> A subject takes p to be relevant to q if and only if, if the subject comes to believe p, then he is disposed to believe q.

We obviously take some propositions to be relevant to others in this sense. For example, someone may take the proposition that my car has been in the parking lot for all day to be relevant to the proposition that I am on vacation. Or one may take the proposition about the date and hour of my birth to be

[7]Dennett (1981a).
[8]Dennett (1981b), 15-6.

relevant to propositions about my character and intelligence. Taking p to be relevant to q should *not* be identified with *believing* that *if* p *then* q (which is interpreted as a material conditional.) If belief is an attitude a subject takes toward propositions, then this relation of taking something to be relevant to another may be regarded as a separate attitude a subject takes toward a *pair* of propositions.

Although, in general, taking something to be relevant to another cannot be identified with beliefs, in some cases, it can be *explained* by beliefs that the subject has, in the following sense:

> A subject's taking p to be relevant to q is *explained* by the subject's belief of s, if and only if anyone who believes that s ought to take p to be relevant to q, in order to be *rational.*

For example, some may take the proposition that my car has been in the parking lot for all day to be relevant to the proposition that I am on vacation *because* he believes that I use my car everyday except when I am on vacation. Anyone who believes the same will take the former to be relevant to the latter. But in general, we don't seem to have a logical guarantee that all such dispositions are explained by beliefs. (We will see some counterexamples later.)

Now I want to define a stance as *a pattern of what considerations someone takes to be relevant to certain propositions.* Let's say that the subject *takes* or *adopts* a certain stance if he is disposed to form beliefs according to the pattern it determines. And we can say that a subject takes a certain stance *toward* (the question whether) p if he takes certain considerations to be relevant to p in conformity to the pattern. This pattern may be *described* in multiple ways. Describing it in terms of "instructions" (as Dennett does) may be one way, or

describing it more formally in terms of some epistemic rules and principles may be another. All of our earlier examples count as stances in this sense. By adopting an astrological stance, one takes propositions about a person's date of birth and hours to be relevant to a person's behaviors in accordance with a specific pattern. Or by taking the intentional stance, one will take propositions about the person's beliefs and desires to be relevant to it in accordance with a specific pattern.

As one's taking some propositions to be relevant to others may be able to be explained by the subject's beliefs in some cases, a stance may also turn out to be explained by beliefs in the same sense. And it seems reasonable to think that many familiar stances can be explained by beliefs in that sense. For example, the astrological stance may be explained by belief in some astrological theory, and the intentional stance by belief in some folk-psychological theory.

Dennett says: "The decision to adopt the intentional stance is free, but the facts about the success or failure of the stance, were one to adopt it, are perfectly objective."[9] Let's say that a stance is successful or reliable when those beliefs formed according to the pattern it defines tend to be true. The astrological stance is not a successful strategy, in that it rarely yields true predictions. Perhaps one stance can be *more* successful or reliable than another in that it tends to yield more true beliefs than the other.

Another important point is this: Two stances may be "rival" strategies. Suppose that you and I are trying to predict a patient's behavior. You take a neurological stance, while I take a psychoanalytic stance. You take some features of the patient's brain to be relevant to it, while I take things like his childhood history to be relevant to it. It may be the case that both stances are success-

[9]Dennett (1981a), 24.

ful stances. The two stances can be said to be "incongrous," in that the sorts of considerations you and I take to be relevant to the same thing is disparate. Moreover, it is not necessarily the case that taking such incongruous stances at the same time makes us contradict each other.

3.5 What Is a Deliberative Stance?

What I want to claim in this section is that there is a distinctive stance (stance precisely in the sense we defined in the previous section) we normally take toward, and only toward, propositions about our own actions. Moran calls this stance a *deliberative stance*, while a stance we take toward other people's actions a *theoretical stance*.[10]

Let's say that one takes a *theoretical stance* toward p if he takes the question whether p to be a matter to be resolved by *evidence* in an intuitive sense. All the examples of stances we saw in the previous sections may count as theoretical stance in this sense. We hardly appeal to evidence to know about our own actions. So this much is clear: We do *not* take a theoretical stance toward questions about our own actions.

Moran persuasively argues that there is a distinctive stance we take toward our own attitudes and actions. His primary concern is with self-knowledge of our own beliefs, and so let us start with the case of beliefs. Consider the following familiar passage from Gareth Evans:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me 'Do you think there is going to be third world war?', I

[10]Although Moran is not always clear on this, Moran also means by stances ways of knowing. He says: "We should ...see [the deliberative stance] and [the theoretical stance] *two ways of coming to know the same thing*"Moran (1997, 154), 154, my emphasis.

> must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p.[11]

The passage describes how we answer questions about our own beliefs. First, consider how some other person will try to answer the question whether you believe there is going to be third world war. He will take some considerations that usually count as evidence to be relevant to the question. For example, he may ask you and see what you say. Or if you are not available, he might try to recollect what you have said on related matters, or your political inclinations, etc. But you, asked the same, that is, whether *you* believe there is going to be third world war, will take those considerations to be rather irrelevant.

Then what considerations do I take to be relevant to such a question? Moran's says:

> [T]he relation of transparency ... concerns a claim about *how* a set of questions is to be answered, what sorts of reasons are to be taken as relevant. The claim, then, is that a first-person present-tense question about one's belief is answered by reference to (or consideration of) the same reasons that would justify an answer to the corresponding question about the world.[12]

That is, the idea is that what you take to be relevant to the question whether you believe p are *reasons* for believing p. So we clearly identify a pattern of what considerations a subject takes as relevant to questions of the form "Do you believe that p. We may describe this pattern in the following way:

(Bel1) If a subject takes p to be relevant to q, then the subject also takes p to be relevant to the proposition that *he (himself) believes that* q.

[11] Evans (1982), 225.
[12] Moran (2001), 62.

Or more simply:

(Bel2) A subject takes p to be relevant to the proposition that *he believes that* p.[13]

Those who are interested in epistemology of self-knowledge of beliefs have further questions to ask. Do those beliefs formed according to this pattern deserve to be called *knowledge*, even though they are not based on evidence? If they are, how? Does this deserve to be called a general model for self-knowledge of beliefs? These may be important questions, but for our purpose, all that matters is that we take such a stance toward the questions of our beliefs.

Something very similar seems to hold for our actions as well. Suppose you are asked, "What will you wear for the party?" Other people, when asked the same question ("What will he wear for the party?"), will try to answer by reference to evidence. For example, what you have wore for that kind of parties recently, whether you believe that celebrities are coming to the party, etc. But you yourself will find those considerations rather irrelevant. The kind of considerations that you will take to be relevant to the question might be: what the dressing code for the occasion is, what kinds of people are coming to it, and what the weather will be like, etc. For example, you may say, "It will rain tomorrow, and it's not a formal occasion. So I will just wear jeans."

It is not hard to notice that these consideration are what the subject takes to be *reasons* for wearing jeans. The relevant notion of reasons here is what are often called "justifying reasons" (as opposed to "explanatory reasons," which are mental states such as beliefs and desires), which are considerations or propo-

[13]Cf. Alex Byrne describes this stance in terms of the epistemic rule of the following form: if p, then believe that you believe p (Byrne 2005, 95) .

sitions that count in favor of doing something. Then the following seems to hold:

(Act) If a subject takes p to be a reason for doing X (or takes p to count in favor of doing X), then he takes p to be relevant to the proposition that *he will do X*.

So we again identify a clear pattern of what considerations one takes to be relevant to propositions about one's own actions. That is, we have a distinctive stance.

We characterized the difference between theoretical and deliberative stances in terms of what *kind* of considerations one takes to be relevant to questions about actions. But there seems to be another significant difference dividing them. Unlike theoretical stances, there seems to be something "subjective" about it, in this sense: I take a deliberative stance only toward propositions about my actions, and you take a deliberative stance only toward propositions about your actions.

3.6 A First Person Perspective *as* a Deliberative Stance

We are ready to put forward our central thesis. I propose the following:

> *The stance view*: Taking a first person perspective on one's actions is nothing but taking a deliberative stance toward those actions.[14]

Taking a third person perspective can be understood in a parallel manner: It is to take a *theoretical stance* toward them. In other words, the difference between

[14]Cf. Moran at one point says, "the agent's perspective is characterized by the dominance of justifying reasons over explanatory ones" (Moran 2001, 131).

the first person and the third person perspectives should be understood as what kind of considerations one takes to be relevant to resolve questions about one's actions. We take evidence to be relevant to the propositions about other people's actions while taking justifying reasons to be relevant to questions about our own actions.

As we saw earlier, some stances can be *explained* by beliefs. We presumed that most of theoretical stances can be explained by beliefs. But if this is really the case, the stance view and the belief view, when applied to the third person perspective, aren't really distinguishable. This is indeed a welcomed consequence; for many would think that taking a third personal point of view is just a matter of beliefs.

But when it comes to the first person perspective, the situation is different. There is reason to think that a deliberative stance cannot be explained by any ordinary beliefs. It's because of what we called *subjectivity* of deliberative stances. I take a deliberative stance toward *my* actions. If this stance could be explained by some beliefs I have, then anyone who has those beliefs would take a deliberative stance toward *my* actions. It is hard to think that there are such beliefs.

Now let us consider how well the stance view can account for the intuitive characteristics of the first personal viewing of one's own actions we listed earlier.

What we called *Incongruity* is the easiest one. A deliberative stance and a theoretical stance are rival stances, in that they take totally different considerations to be relevant to one and the same questions. I think this captures the sense that they are incongruous. Nonetheless, they both are successful stances, tending to yield true predictions in this particular case. They are not contradic-

tory with each other in any sense.

By *Subjectivity.* The belief view had troubles with this, because this seemed to imply that there are private beliefs, only accessible to particular persons. But there is no corresponding problem here. Moreover, what we called subjectivity of deliberative stance seems to match *Subjectivity.* We take a deliberative stance only toward our own actions.

Rationality says that taking a first person point of view toward one's impending actions is a rational requirement. We can at least see that a deliberative stance has a similar property. It seems that there is something wrong with a subject who doesn't take a deliberative stance toward his own actions. But as for *why* a deliberative stance has this feature, we don't have any explanation yet.

What about *Belief-likeness*? We said earlier that this feature strongly supports the Belief View, and that it may be even doubtful whether there is any alternative account that can explain it. Nothing we have said so far suggests that taking a deliberative commits oneself to a conception of how one oneself is related to some objective states of the world.

So two questions remain: How can taking a deliberative stance itself be regarded as committing the subject to a self-conception? How can to take a deliberative stance be a rational requirement?

3.7 How a Deliberative Stance Makes Up a "Self-Conception"

When we view ourselves from internal perspectives, we view ourselves as "the authors of our actions and the leaders of our lives." We seem to be committed to a certain *conception* about how we ourselves related to our actions. It is

about the source of actions. But what we have here is at best an account of the way we *know* about them. Didn't we simply change the subject? But I want to claim that a deliberative stance itself may be regarded as embodying a certain *conception* about how the subject, *qua* the subject of conceptions, is related to some objective states of affairs. What I will say here will be somewhat speculative and sketchy, but I hope it will lessen the worry a little.

Let me start with the following passage from Wittgenstein's *Tractatus*.

> If I wrote a book "The world as I found it," I should also have therein to report on my body and say which members obey my will and which do not, etc. This then would be a method of isolating the subject or rather of showing that in an important sense there is no subject: that is to say, of it alone in this book mention could not be made.[15]

Suppose that I am trying to write such a book. It purports to contain every minute detail of the world as I find it. (Let us suppose that I am supposed to use only objective languages in it; in particular, I am not supposed to use any indexical expressions in it.) As the world as I find it contains numerous people (including the person who I take to be myself), the book will feature many people and all the detailed information I believe about them. In order for this book to contain all the information about the world as I find it, don't I have to indicate *in* the book that one of those people is *myself*, that is, the author of the very book, and that one among those numerous people "obeys my will"—that is, I am the originator of a certain objective person's actions? How can I achieve that? Perhaps if person X is the one I take to be myself, I want to write "X obeys X's will." But obviously, that won't do, unless I manage to indicate that X is myself. I may be tempted to say "X is the author of *this* book," but

[15]Wittgenstein (1922), 5.631.

you are not allowed to do that. So Wittgenstein seems quite right in saying that "of it alone in this book mention could not be made."

Now I think that we can see here how two different *perspectives* are generated. First, I may ask how certain person X is related to behaviors of X. But I may also ask how I, the author of the book, is related to behaviors of X who is in the book. The first amounts to occupying the third personal point of view toward X, while the second amounts to taking a first personal point of view toward X. The answers to the first question can be easily represented inside book; it will be mostly about causal stories of how X's mental states cause X's behaviors, etc. But the answers to the second question can't be represented inside the book.

Our belief is analogous to such a book. Belief purports to represent the objective world as it is in itself, and we represent it as propositions believed, in abstraction of the subject of belief or the state of believing, where propositions are absolutely true or false (just as a book written only with objective languages). So propositions are the analogue of the book, and the subject of belief is the analogue of the author of the book. Just as who in the book obeys the author's will, and who in the book is the author cannot be written inside the book, who in his belief obeys the will of the subject of the belief, and who in his belief is the subject of the belief cannot be represented in beliefs. (And that's why the belief view is bound to be false.)

But all this should sound paradoxical. Don't I obviously have the ability to represent that I myself, as the very subject who is representing, have control over behaviors of someone? So we seem to be able to achieve something that should be impossible in principle. But how? I want to claim that something like a *deliberative stance*, which cannot be explained by belief, provides a round-

about but effective means to "represent" how the subject of representation is related to something that is represented.

Returning to the book analogy, let's ask this question. What other resources do I, as the author of the book, need to have in order to make the readers of the book to be able to tell which of the people featured in the book is the author of the book, or which "obeys my will"? Here is one way: Suppose that I write revised editions of the book, as the world as I find it will be continuously updated. And suppose that I can make available to the readers of the book how the subsequent editions of the book will be developed, upon various alternative scenarios. (That is, the readers know how the book is disposed to be revised under alternative situations.) The readers may identify something like the following pattern: if the 2nd edition of the book contains "p," then the 3rd edition of the book contains "X believes that p". Or if the 2nd edition of the book contains "it is raining," then the 3rd edition contains "X will bring the umbrella." The readers, finding such an unusual pattern of revision, will have an almost foolproof method to tell who the author is, and moreover, how the author's view of the world is related to X's behaviors.

And I think that's precisely how we can achieve something that seemed impossible. Taking a deliberative stance is a matter of how one's belief is disposed to be revised, and it is capable of containing the information about how a certain objective person's behaviors is responsive to the subject's own conception of the world. So I want to claim that a deliberative stance itself can be justly regarded as being constitutive of *self-conceptions* one has reflexively about oneself. That is, by taking a deliberative stance toward a certain person's actions, I am committing myself to a view of how that person's actions are influenced by

my own conception of the world.[16]

If something is to be representational, then it should be true or false. We can't have precisely that, but instead substitute notions. A stance can be reliable or not, and individual applications of it can be correct or incorrect. But what makes it right is not some mysterious "subjective facts," but plain objective facts.

3.8 A Deliberative Stance and Rationality

The last thing we want to consider is the question why taking a deliberative stance toward one's actions is so much as a *rational requirement* for the agent.

The conception of rationality I want to use is a very basic one, which doesn't involve the notion of reasons at all. According to this conception, an action is rational when it rationalized by the agent's belief and desires. And then the claim comes down to this: the agent's taking a deliberative stance toward his own actions is a necessary condition for those actions to be rational. How can this be right? In order for an action to be rationalized, isn't it enough for the agent to have adequate beliefs and desires? But then what remains to be done by anything like a deliberative stance?

Let us look at a very elementary picture of rational actions.

> What is essential to rational action is that the agent be confronted, or conceive of himself as confronted, with a range of alternative possible outcomes of some alternative possible actions. The agent has attitudes, pro and con, toward the different possible outcomes, and

[16]Very notable in this context is that it is generally agreed that there is something subjective with *conditional beliefs*. Some think that this feature makes them defective. But I think we should look at a bright side of it. For it provides a means to achieve something that "non-defective discourse" can't. Conditional beliefs can be *reflexively* "about" the subject's belief state. Moreover it seems to be an only means to achieve that. See e.g. Stalnaker (1984), Ch. 6 and Bennett (2003).

> beliefs about the contribution which the alternative actions would make to determining the outcome. One explains why an agent tends to act in the way he does in terms of such beliefs and attitudes.[17]

It is right that beliefs and desires completely determine what it is a rational thing for the agent to do. But it presupposes that "the agent [is] confronted, or conceive himself as confronted, with a range of alternative outcomes of some alternative possible actions." We earlier called these alternatives that "seem to lie open before an agent" options for the agent, and we saw this determines a partition of the space of the agent's doxastic possibilities. What determines such a partition? We have an obvious answer in hand: If you take a deliberative stance toward p, then p is an option for you. If this is right, then for any action to be rationalizable, we need to assume that the agent takes a deliberative stance toward one's own actions.

[17]Stalnaker (1984), 4.

Bibliography

Anscombe, G. E. M. (1957). *Intention.* Cambridge, MA: Harvard University Press, 2nd edition.

Armstrong, D. M. (1973). *Belief, Truth and Knowledge.* Cambridge University Press.

Bennett, J. (2003). *A Philosophical Guide to Conditionals.* Oxford University Press.

Byrne, A. (2005). Introspection. *Philosophical Topics* 33:79–104.

Castañeda, H. N. (1968). On the Logic of Attributions of Self-Knowledge to Others. *The Journal of Philosophy* 65:439–456.

Dennett, D. C. (1981a). True Believers: The Intentional Strategy and Why It Works. In Heath, A. F., ed., *Scientific Explanation.* Oxford University Press. Reprinted in D. C. Dennett, *The Intentional Stance*, 1987, MIT Press.

— (1981b). Where Am I? In *Brainstorms.* Cambridge, MA: MIT Press.

— (1984). *Elbow Room: The Varieties of Free Will Worth Wanting.* The MIT Press.

— (1992). *Consciousness Explained.* Back Bay Books.

DeRose, K. (1991). Epistemic Possibilities. *The Philosophical Review* 100:581–605.

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

Frege, G. (1919/1997). Thought. Reprinted in M. Beaney ed. *The Frege Reader.* Blackwell Publishers.

Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly* 32:127–136.

Jeffrey, R. C. (1990). *The Logic of Decision.* University of Chicago Press, 2nd edition.

Kaplan, D. (1989). Demonstratives. In Almog, J., Perry, J., and Wettstein, H., eds., *Themes from Kaplan.* Oxford University Press.

Korsgaard, C. M. (1989). Personal Identity and the Unity of Agency: A Kantian Response to Parfit. *Philosophy & Public Affairs* 18:101-132.

Lewis, D. K. (1979). Attitudes *De Dicto* and *De Se. The Philosophical Review* 88:513-45. Reprinted in Lewis (1983c).

— (1981). Causal Decision Theory. *Australasian Journal of Philosophy* 59:5-30. Reprinted in Lewis (1986).

— (1983a). Individuation by Acquaintance and by Stipulation. *The Philosophical Review* 92:3-32. Reprinted in Lewis (1999).

— (1983b). New Work for a Theory of Universals. *The Australasian Journal of Philosophy* 61:343-377. Reprinted in Lewis (1999).

— (1983c). *Philosophical Papers*, volume I. New York: Oxford University Press.

— (1986). *Philosophical Papers*, volume II. New York: Oxford University Press.

— (1988). What Experience Teaches. *Proceedings of the Russellian Society* 13:29-57. Reprinted in Lewis (1999).

— (1994). Reduction of Mind. In Guttenplan, S., ed., *A Companion to the Philosophy of Mind.* Basil Blackwell. Reprinted in Lewis (1999).

— (1999). *Papers in Metaphysics and Epistemology.* Cambridge University Press.

Lycan, W. B. (1995). A Limited Defense of Phenomenal Information. In Metzinger, T., ed., *Conscious Experience.* Exeter: Imprint Academic.

Moore, G. E. (1944/1993). Moore's Paradox. In Baldwin, T., ed., *G. E. Moore: Selected Writings.* London: Routledge.

Moran, R. (1997). Self-Knowledge: Discovery, Resolution, and Undoing. *European Journal of Philosophy* 5:141-161.

— (2001). *Authority and Estrangement: An Essay on Self-Knowledge.* Princeton, NJ: Princeton University Press.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review* 83:435-450.

— (1983). The Objective Self. In Ginet, C. and Shoemaker, S., eds., *Knowledge and Mind.* Oxford University Press.

— (1986). *The View from Nowhere.* New York: Oxford University Press.

Nida-Rümelin, M. (1995). What Mary Couldn't Know: Belief about Phenomenal States. In Metzinger, T., ed., *Conscious Experience.* Exeter: Imprint Academic.

Ninan, D. (2009). Persistence and the First-Person Perspective. *The Philosophical Review* 118:425–464.

Perry, J. (1977). Frege on Demonstratives. *The Philosophical Review* 86:474–497. Reprinted in Perry (2000).

— (1979). The Problem of the Essential Indexical. *Noûs* 13:3–21. Reprinted in Perry (2000).

— (2000). *The Problem of the Essential Indexical and Other Essays.* Stanford, CA: CSLI Publications, expanded edition.

— (2003). *Knowledge, Possibility, and Consciousness.* Cambridge, MA: The MIT Press.

— (2006). Stalnaker and Indexical Belief. In Thomson, J. and Byrne, A., eds., *Content and Modality: Themes from the Philosophy of Robert Stalnaker.* Oxford University Press.

Savage, L. J. (1972). *The Foundations of Statistics.* Dover Publications.

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

— (1995). Moore's Paradox and Self-Knowledge. *Philosophical Studies* 77:211–228. Reprinted in Shoemaker (1996).

— (1996). *The First-Person Perspective and Other Essays.* Cambridge University Press.

Stalnaker, R. C. (1981). Indexical belief. *Synthese* 49:129–151. Reprinted in Stalnaker (1999a).

— (1984). *Inquiry.* Cambridge, MA: MIT Press.

— (1999a). *Context and Content.* New York: Oxford University Press.

— (1999b). Extensive and Strategic Forms: Games and Models for Games. *Research in Economics* 53:293–319.

— (2003a). On Thomas Nagel's Objective Self. In Stalnaker (2003b).

— (2003b). *Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays.* Oxford University Press.

— (2006). Response (to Perry, "Stalnaker and Indexical Belief"). In Thomson, J. and Byrne, A., eds., *Content and Modality: Themes from the Philosophy of Robert Stalnaker.* Oxford University Press.

— (2008). *Our Knowledge of the Internal World.* Oxford University Press.

Velleman, J. D. (1989). Epistemic Freedom. *Pacific Philosophical Quarterly* 70:73–97. Reprinted in Velleman (2000).

— (1996). Self to Self. *The Philosophical Review* 105:39–76. Reprinted in Velleman (2006).

— (2000). *The Possibility of Practical Reason.* New York: Oxford University Press.

— (2006). *Self to Self: Selected Essays.* Cambridge University Press.

Williams, B. (1972). Imagination and the Self. In *Problems of the Self.* Cambridge University Press.

Williamson, T. (2002). *Knowledge and Its Limits.* Oxford University Press.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus.* Translated by C. K. Ogden, New York: Routledge and Kegan Paul.

— (1958). *The Blue and Brown Books.* Blackwell Oxford.

— (1974). *Philosophical Investigations.* Trans. G. E. M. Anscombe. Oxford: Blackwell.