# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**UMI**

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

# EXTERNALISM AND SELF-KNOWLEDGE

by

ADAM S. VINUEZA

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

1996

**This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.**

August 7, 1996
**Date**

**Chair of Examining Committee**

August 7, 1996
**Date**

**Executive Officer**

**Supervisory Committee**

**THE CITY UNIVERSITY OF NEW YORK**

**1996**

## Abstract

**EXTERNALISM AND SELF-KNOWLEDGE**

**by**

**Adam S. Vinueza**

**Advisor: Professor Stephen Schiffer**

**It is often alleged that semantic externalism, the thesis that mental content is partially determined by external factors, conflicts with the view that we have some kind of privileged access to the contents of our own mental states. If this is correct, we have a paradox: externalism is powerfully intuitive, but so is the idea of privileged access. In this essay, I examine the apparent paradox and argue that the alleged conflict is illusory.**

**After sketching the paradox and its significance in Chapter I, I take a close look at privileged access (II) and externalism (III).**

**In Chapter IV, I address arguments for the incompatibility. I argue that all the significant arguments for it (due to Paul Boghossian, Anthony Brueckner, and Michael McKinsey respectively) turn out to be the same**

argument; I call it the Improved McKinseyan Argument (IMA). In essence, IMA alleges that it is possible to infer from non-empirically justified (hence, privileged) knowledge of one's own thoughts and of externalism that certain paradigmatic empirical facts obtain; because it seems that one can know in a privileged way what one deduces from other privileged knowledge, it seems to follow, contrary to intuition, that one can have privileged knowledge of empirical facts.

In Chapter V, I address the significant solutions offered by others to the paradox, and argue that none of them work.

I turn in Chapter VI to a fresh examination of IMA. Crucial to the argument is the notion of non-empiricality it appeals to, and I argue that it divides into three separate notions. It turns out that on no consistent understanding of non-empiricality are all the premises of IMA true. I then diagnose the motivation for IMA, that is, what made it seem compelling in the first place. I suggest that two factors are at work here: first, reliance on an unexplicated notion of non-empiricality, and reliance on an unexplicated and controversial notion of privileged access.

Chapter VII briefly addresses the underlying explanation for privileged access, and in it I suggest that what explains its non-empirical authority is a combination of naturalistic and conceptual factors.

## ACKNOWLEDGEMENTS

My first thanks are due to Stephen Schiffer for all his encouragement and help. As will become apparent in the chapters which follow, my work has been strongly influenced by his. I am also grateful to Hartry Field and David Rosenthal for much help and useful criticism. Hartry's criticisms and comments helped shaped my views about defeasibility and the a priori. David has taught me just about everything I know about consciousness, and I made use of some of this knowledge in the chapters on privileged access. (He is of course not responsible for my errors.)

Thanks also to my friends Jody Azzouni, Jonathan Adler, and Peter Ross, who were all very generous with their time and criticism in discussions of this material.

Especially warm thanks are due to my friends Celeste Friend and Maureen Linker, whose friendship, counsel and support are an important reason why this dissertation is finished and is as good as it is.

Primordial versions of Chapter VI were read at the University of Colorado at Boulder, The Ohio State University, and the University of Toledo. I am grateful to the audiences at each university for their comments, but

especially to George Bealer, Stephen Leeds, Diana Raffman, and Bill Taschek.

Finally and most importantly, I thank my wife, Lisa Bates, whose unqualified love and support have been and always will be invaluable to me. Without Lisa, this dissertation simply would not have been.

## DEDICATION

To my father, Silvio Vinueza, and to the memory of my mother, Leona Schwartz Vinueza.

# Table of Contents

# I

# INTRODUCTION TO THE PARADOX

In this essay, I am concerned with a seemingly paradoxical consequence of semantic externalism: that it somehow conflicts with the view that we have privileged access to the contents of our own mental states. I will argue that the seemingly paradoxical consequence is a mere seeming, and that everything, semantically and epistemically speaking, is all right. In this brief chapter, though, I will only give the flavor of the alleged paradox, say a little about how best to approach its solution, and sketch the rest of the essay.

## 1. Preliminary: twin-earth thought experiments

Externalism is motivated by the well known thought experiments of Putnam (1975) and Burge (1979). Putnam hypothesizes a 'twin' earth which is identical to earth but for the presence on twin earth of a substance distinct from $H_2O$ which plays the role that water plays on earth. For every person on earth, then, there will be a molecule-for-

molecule 'twin' on twin earth. (Let's call the residents of twin earth 'twearthlings' and the local waterish substance 'twater'.) Consider a pair of such twins. They will be alike in every respect save one: the earthling's concept associated with the word 'water' will be about water, while the twearthling's corresponding concept will be about twater. That's because whatever reasons we have for ascribing beliefs about water to earthlings who utter sentences such as 'A glass of cold water on a hot day sure is refreshing' and 'There is not much water in the Sahara desert' are reasons for ascribing beliefs about twater to twearthlings who utter type-identical sentences.

Burge's hypothesis is rather different. He doesn't hypothesize a twin to earth, but let's stay with it for ease of exposition. Suppose that twin earth also differs from earth in the following respect: while the word 'arthritis' refers in English only to joint ailments, the word 'arthritis' is used by twearthlings (who speak Twenglish) to refer to both joint and bone ailments. A competent speaker of English may well falsely assent to the sentence 'I have arthritis in my thigh'--because she believes (falsely) that

she can have arthritis in her thigh; that speaker's twin, however, would (by Twenglish standards) correctly assent to the very same sentence. Because arthritis cannot be had in one's thigh, we can't explain the fact that her twin would truly assent to the sentence by appeal to her believing truly that she can have arthritis in her thigh; instead, we must suppose that what she believes when she asserts 'I have arthritis in my thigh' is a proposition about what me might as well call 'twarthritis'--the ailment that plays the role for the Twenglish speaker that arthritis plays for her twin. That is, she does not have arthritis beliefs, even though she is molecule-for-molecule identical to someone who does.

There is some controversy over what these thought experiments actually show. Most theorists, though, have taken them to show that a thinker's propositional attitude state having the content it has depends on her bearing causal relations of some sort to features of her environment or linguistic community. As we'll see in Chapter 3, demonstrating this is far from simple, and it's arguable that only a much weaker thesis strictly follows from these thought experiments alone; we must adduce additional

considerations to motivate those versions of externalism which are popular today.

For the moment, though, let's suppose that the thought experiments do show that thinking a thought depends on bearing causal relations to one's environment or linguistic community. That way, it will be easier to see how the paradox might arise.

## 2. Preliminary: privileged access

When I know what I think or feel, I don't seem to have or need evidence of any kind; but when I know what you think or feel, my knowing clearly depends on my having evidence of some sort. What's more, when I know my experiences, I seem to do so without having to draw inferences; but I can't know about the experiences of others without drawing inferences from my evidence.

This asymmetry points to a strongly intuitive feature of self-knowledge: that our access to our own mental states is privileged, by virtue of being immediate (that is, non-inferential) and authoritative in a way that doesn't depend on empirical evidence. Traditionally, privileged access was held to be virtually definitive of mentality: mental states

just were those things to which we had privileged access. Anecdotal evidence about non-conscious mental states and empirical evidence from such fields as cognitive science, however, have threatened the notion of privileged access by showing how one might have false beliefs about one's own mental states, or no beliefs about them at all.

In Chapter II, I defend privileged access against these threats, by offering a definition of it which plausibly rules out such troublesome phenomena. But at the moment, what's relevant is that privileged access, if it exists, is non-inferential and non-empirically authoritative knowledge of our own mental states. It's quite plausible that we have privileged access to many of our mental states, and it's this that helps lead us to paradox.

## 3. Arguments for the conflict

We are now ready to see how the conflict between externalism and privileged access might arise. If environmental and social facts partly determine mental content, then it can be hard to see how we can have non-empirically authoritative knowledge of our own mental states, for the following two reasons. First, knowledge of

content seems to require knowledge of the facts that determine content. We cannot justifiably ascribe a belief about water to someone unless we are in a position to say that their beliefs are about water, which would seem to require knowledge of facts about water. If this holds for others, why should it not also hold of ourselves? Second, these facts seem to be justifiable through empirical observation and inference alone.

These considerations lead to two styles of argument that externalism and privileged access conflict. The first is skeptical: according to this strategy, it is argued that because we can't know without empirical observation that we bear the appropriate causal relations to the environment or community, we wouldn't know what we are thinking without empirical observation. But we clearly do know what we are thinking without empirical observation, because we have privileged access to our own thoughts. Therefore, externalism and privileged access are incompatible. This popular strategy has been taken by Paul Boghossian (1989) and Anthony Brueckner (1990).

The second strategy is a reductio: according to this strategy, if externalism were true and we had privileged access, then because externalism is knowable in a non-empirically authoritative way, we should be able to infer in a non-empirically authoritative way that we bear such-and-such causal relations to worldly features. But we clearly can't know such things in a non-empirically authoritative way: such knowledge is paradigmatically empirical, hence empirically defeasible. Because externalism and privileged access together lead inexorably to this contradiction, the two are incompatible. This strategy has been taken by Michael McKinsey (1991).

But whatever the argument, a paradox arises, which consists of the following three propositions:

(1) Externalism is true.

(2) We have privileged access.

(3) (1) and (2) are incompatible.

We've already seen that each of (1)-(3) are independently plausible, and it's obvious that the three propositions are mutually inconsistent. Therefore, we must reject at least one of (1)-(3).

## 4. The significance of the paradox

Some theorists have taken the conflict between externalism and privileged access to be a reductio of externalism. (See, e.g., Laurence Bonjour 1991.) Some others take it to be a reductio of the idea that we have privileged access. (See Andrew Woodfield's introduction to his 1982). Most theorists, though, take it to be a serious problem, and try to show that the conflict is illusory.

Obviously, those who reject (1) or (2) are suspicious of either externalism or privileged access, while those who reject (3) believe in both. But what's so important about who's right here?

We want to know who's right because both (1) and (2) have very strong intuitive plausibility. It is by no means unheard of to deny externalism or privileged access, but as will become clear later on, the intuitions supporting them are so strong and unambiguous that denying them solely because there are arguments that they conflict in and of itself incurs upon one the burden to examine these arguments very carefully. Perhaps, when all is said and done, externalism will turn out to be false; perhaps, too, we will

eventually recognize that we do not really have privileged access to the contents of our own mental states. But merely recognizing that the two theses conflict is hardly a sufficient reason for rejecting either.

## 5. Responding to the paradox

It's natural, then, that by way of responding to the paradox, many have tried to refute (3), the claim that externalism and privileged access are incompatible. It's unfortunate that few of them have addressed the best arguments supporting it, but this is understandable; until very recently, there simply were no detailed arguments for the incompatibility, and those who sensed its possibility had to construct arguments themselves--arguments whose conclusions they already knew they disbelieved. For example, two staunch externalists, Donald Davidson (1987) and Burge (1988), argued separately in now famous papers that externalism in no way conflicts with privileged access, but neither found any compelling reasons for thinking that such a conflict exists; and when prima facie compelling arguments appeared, it turned out that the considerations raised by Davidson and Burge lacked the resources to refute

them. I discuss these and other responses to the paradox in Chapter V.

The strategy I take in this essay is to carefully examine each of the best arguments for each proposition making up the paradox, and ask of each proposition whether we should reject it. I discuss (2) in Chapter II, and (1) in Chapter III; perhaps unsurprisingly, I defend (1) and (2) and conclude that the guilty proposition is (3). In Chapter IV, then, I examine the best arguments for (3); I conclude that only one is even prima facie sound, but leave until later to explain why even this best argument is unsound.

Still, showing that the best argument fails is no easy task. As we'll see in Chapter VI, McKinsey's argument, which I take to be the best, is very nearly sound. To explain why it fails in the end, we'll need to closely examine the notion of non-empiricality.

What's more, even if we grant that we can refute the best arguments for (3), we can't at that point claim to have resolved the paradox. To truly resolve a paradox, it is not enough for us to show that one of the propositions making it up is false; we must also explain why it seemed to be true.

This entails explaining why the best argument for the paradox seemed sound, despite being unsound: that is, why it was compelling in the first place. At the end of Chapter VI, I offer such an explanation.

## 6. Explaining privileged access

The considerations I raise to resolve the paradox of externalism and self-knowledge do not depend on any contentious claims about privileged access. Nevertheless, given that I defend privileged access in an intuitive but admittedly hand-wavy way in Chapter II, I ought to answer the question of whether we can offer an informative explanation of the phenomenon of privileged access.

I'll have little to say about what explains privileged access, but I'll suggest that any explanation ought to be relational, that is, in terms of an extrinsic relation between higher-order beliefs and lower-order ones. An explanation of the phenomenon of consciousness has been offered by David Rosenthal (1986 and others), and I make a brief appeal to this explanation as a way of accounting for the non-inferentiality of privileged access. Its non-empirical authority, though, is another matter entirely; I

can't discuss the topic in any depth, but will sketch some of the main issues involved.

## 7. Segue

We now have a good idea of what the paradox of externalism and self-knowledge is, and how I plan to examine it. Our first order of business, then, is to motivate each of the premises of the paradox. This work begins in the next chapter with the notion of privileged access.

# II

# PRIVILEGED ACCESS

## 1. Introduction

It's a truism that there's considerable difference between knowing about my own mental states and knowing about those of others. When I know what I think or feel, I don't seem to have or need evidence of any kind; but when I know what you think or feel, my knowing depends on my having evidence of some sort. What's more, when I know my experiences, I need draw no inferences; but I can't know about the experiences of others without drawing inferences from my evidence. What explains the difference?

An interesting question, but not one I'll answer in this chapter. Here I'll merely set out a view of privileged access I think we can all agree on, then show how it says enough about privileged access so that we can see how it and externalism might conflict. I'll return to privileged access in Chapter VII, and say a little about what explains it there.

What needs to be explained about privileged access? Because it's a form of knowledge, privileged access has both psychological and epistemic features, and these features call for explanation. I have little of substance to say in this chapter about either of these features for a simple reason: I needn't to motivate the paradox.

Much of this chapter is devoted to privileged access's epistemic features. To understand why, in having privileged access, we have a kind of non-empirical authority of our own mental states, we should examine its epistemic features. By contrast, little of this chapter is about psychological features, but we'll examine it some more in Chapter VII.

## 2. Psychological features

### 2a. Non-inferential higher-order belief

I've already pointed to one psychological feature of privileged access: that coming to know about one's own present experiences doesn't seem to involve inference. But one can't know about one's own experiences without being aware of them somehow. How, then, are we aware of them?

The obvious, uninformative answer of tradition is that we introspect them. What philosophers have called introspection just is that awareness of one's mental states by virtue of which one knows about them in a privileged way[1], so to appeal to introspection without further elaboration is just to redescribe what we wish to explain. The same holds for the view that introspection is some kind of "inner sense".[2] Inner sense certainly isn't perception, so it must be like perception in some relevant way. What way is that?

We can avoid these detours if we keep in mind that privileged access is a kind of knowledge, and that to have

---

[1] I'm avoiding some subtleties here. The ordinary, everyday awareness we have of our mental states doesn't really deserve to be called introspection, even though we come to know about our mental states most immediately in this way. Introspection, as we commonly understand it, is more deliberate and focused. (For a nice discussion of these differences, see David Rosenthal (1986), esp. section IV.) Nothing here rides on such subtleties, however.

[2] This is not to say that "inner sense" views are vacuous, but only that privileged access is explained only to the extent that inner sense is explained; the appeal to inner sense without further elaboration is unhelpful.
Sydney Shoemaker has discussed the "inner sense" view of introspection at great length in the 1993 Josiah Royce Lectures (Shoemaker 1993).

knowledge about one's mental states is, at least, to have beliefs about them. (It's implausible to think that this knowledge is only know-how. Knowing how to, say, detect and discriminate one's mental states comes with knowing that one is in the detected states and not others.) Beliefs are much less mysterious than inner sensings--we know what beliefs are, more or less, but we have no idea what inner sensings are. Let's suppose, then, that when we have privileged access to our mental states, we have beliefs about them (that is, higher-order beliefs--beliefs about mental states, as opposed to beliefs about non-mental things), and that these beliefs seem to be arrived at non-inferentially.[3]

2b. Self-knowledge differs only with respect to its object

The other relevant psychological feature of privileged access is that the difference between knowledge of one's sensations and knowledge of one's thoughts seem to differ only to the extent that sensations and thoughts differ from one another. We don't seem to have different ways of

---

[3] It's worth noting that these beliefs may well be arrived at inferentially, so long as the inferences involved aren't ones we're conscious of making. For an elaboration of this point, see Rosenthal (1986).

knowing about them, whatever that would mean. This constrains any explanation of privileged access, because any explanation which held for thoughts and not sensations (or vice-versa) would have to also explain why we only seem to know our thoughts and our sensations in the same way.

## 3. Epistemic features

### 3a. Privileged access as self-warranted belief

I claimed in the previous section that privileged access seems non-inferential. This may make it seem as if justification is irrelevant, especially if one thinks that the only way a belief can be justified is by being inferred from evidence. But it's not obvious that this way of thinking about justification is correct. We tend to think of ordinary perceptual beliefs as being justified, and these beliefs aren't in general arrived at through inferences from evidence. It's possible to explain this by claiming that what justifies ordinary perceptual beliefs is their reliability: that they are the results of reliable belief-forming processes, or something like that. But one could equally well hold that even perceptual beliefs are justified

by virtue of there being evidence for their truth, even though the perceiver needn't have inferred the belief from this evidence. No: one could hold that one simply has the ability to ground these beliefs by inferring them from one's perceptual evidence, and thereby count those beliefs as justified.

I must admit that I'm skeptical of the latter view of justification, but it doesn't matter when it comes to privileged access, because it's hard to see just what evidence one might have that would ground one's beliefs about one's own mental states. What sort of evidence would this be? In the case of perceptual beliefs, we can point to other perceptual beliefs as evidence, or perhaps to beliefs about our own perceptual states; but we have no idea what sort of available belief or experience would count as evidence for what we believe about our own mental states. When I believe that I'm thinking, say, that water quenches thirst, I am utterly at a loss as to what would count as evidence for my belief. At best, I could say only that I just know that I'm thinking that thought--that I simply have no reasons for believing it.

Contrast this phenomenon with perceptual beliefs. If someone asks me what reasons justify my visually-formed belief that an elephant is walking down Fifth Avenue, I can say that I believe that I'm seeing an elephant walking down Fifth Avenue, that my eyesight is good, that I know what elephants are, etc. Reasons for my belief fall trippingly from my tongue. This does not happen when I'm asked what reasons justify my belief that I'm thinking that an elephant is walking down Fifth Avenue. That my eyesight is good justifies my thought, not my belief about it; the same holds for my knowledge about elephants, and so it seems for any other considerations that might come to mind. It would be ludicrous for me to justify my belief with gerrymandered analogous kinds of evidence: e.g., that I believe that I'm thinking that thought, that my introspective faculty seems to be in working order, that in the past whenever I've believed I've had a thought I've been right, and so on.

This suggests that my simply believing that I'm in some mental state is warranted independently of any reasons anyone might have for thinking that I'm in that state. One way of describing this feature is by saying that privileged

access is "self-warranted," in that having the relevant belief implies that the belief is warranted. William Alston expresses the idea as follows:

> Each person is so related to propositions ascribing current mental states to himself that it is logically impossible both for him to believe that such a proposition is true and not be justified in holding this belief.[4]

Set aside the part about logical impossibility, which is a bit dubious. Alston is claiming that there is no way to believe that one is in some mental state without being justified in doing so.

Is Alston right? I don't know, but it seems to me that he is relying on a very important intuition about self-knowledge to bolster his claim. The intuition which justifies Alston's claim that we can't believe we're in a mental state without being justified in believing so is that we have no idea what it would be like to believe we're in a mental state without being justified in believing so. This is an important intuition, and although it won't be relevant here, we'll come back to it when explaining privileged access in Chapter VII. For the moment, though, we should

---

[4] Alston (1971), 235.

take a look at some phenomena which seem to straightforwardly refute the view that the epistemic privilege of privileged access is adequately captured by Alston's claim that beliefs about our own mental states are self-warranted.

### 3b. Non-conscious phenomena and introspective fallibility

Many mental states are states we're not at all conscious of being in. This quickly leads to trouble for Alston's notion of self-warrant, for an obvious reason: we can know about utterly non-conscious mental states, but we can't do so introspectively. This means that we must reason our way to beliefs that we are in such states. Surely, these beliefs are not self-warranted, but instead are much more like beliefs we have about the mental states of others. We draw inferences about what non-conscious mental states we're in on the basis of sophisticated theories of our own behavior and its psychological significance.

For example, English speakers will understand the sentence, "John went to the bank" as being ambiguous between John's having gone to the side of a river and his having gone to a financial institution. But it's been shown that

after a word semantically related to "river" (say, "ocean") is shown to them just beneath their visual detection threshold (say, flashed on a screen for a tenth of a second), they tend to interpret the sentence as being about a river and not about a financial institution. This and many similar experiments show that people can process semantic information unconsciously; we must therefore be able represent the meanings of words without being aware of doing so.

Now, the kind of reasoning I offered in the last paragraph could be intended to show that some person mentally represented the meaning of "ocean" without realizing it. I can even apply such reasoning to myself. But in that case, my warrant for believing that I was in a mental state which represented the meaning of "ocean" does not lie simply in my having that belief; it lies mainly in there being evidence to that effect, and a theory positing that belief which best explains the occurrence of that evidence. It's easy to see, then, that if Alston's claim were correct, there couldn't be any non-conscious mental states, because such states would be just the sorts of

states we couldn't have self-warranted beliefs about. But there is strong evidence that there are non-conscious mental states.

Another problem for Alston's characterization of privileged access is the fact that introspection is fallible in various ways. Here's one way: I might be in the shower when suddenly the temperature of the water changes drastically. For a few moments, I'm not sure whether I'm experiencing extreme hot or extreme cold. I know that I'm experiencing one or the other, I just don't know which.[5] Here's another way: because of subsequent experience, I may not accurately remember my experiences at some previous time. I might think that I saw someone wearing a tie who wasn't, for example. Indeed, I might vividly seem to

---

[5] The identification theorist might respond that until I realize what I'm experiencing, I'm not experiencing either extreme hot or extreme cold, but instead something else--say, an experience as of something that's either extremely hot or extremely cold. But this is hard to believe, and doesn't accord with the way we talk about sensations. If I soon realize that the water has turned extremely hot, I wouldn't say that I first had a neutral experience and then had a different one of extreme hot; I would say that I had a single experience of extreme hot from the very beginning, but that I wasn't sure that it was such an experience until later on.

remember having seen that person wearing a tie, when I had no such experience.[6] Here's a third way: in a taste test of four dishes of apple sauce laid out from left to right on a table, I might judge the fourth dish I tasted--the rightmost--to taste the best. When asked for an explanation of why I thought it tasted best, I might say that its taste was in some special way different--that it was the smoothest, perhaps. As it happens, all four dishes of apple sauce are from the same jar, and the reason I preferred the last one has to do with the facts that English speakers read from left to right and that people have a marked preference for things seen most recently, other things being equal.[7]

These ways of being introspectively fallible strongly suggest that even when we have a non-inferential belief about one of our own conscious mental states, this belief might not be warranted simply by virtue of our having it. In the shower case, I might rashly judge that I am feeling sensations of extreme heat, when I'm not; in the case of the

---

[6] This phenomenon has been discussed extensively by Daniel Dennett (1991).

[7] This is an instance of what Nisbett and Wilson (1977) call the "position effect."

non-existent tie-bearer, I have a false belief about one of my past experiences; in the apple-sauce case, I am making up a story about my reasoning process which isn't in the least bit true.

Each of these cases involves unwarranted belief. In the respective special cases, there is no good reason for thinking I'm warranted in believing that I am having sensations of extreme heat, or that I recently saw such-and-such a man wearing a tie, or that I tasted something special in the rightmost dish of apple sauce. Not only are there explanations for why such beliefs are unreliable, these explanations are also available to me; I can appeal to them to explain why my own beliefs might well be mistaken. Because it's not plausible to claim that such beliefs are warranted despite my knowing that they're unreliable, we can only conclude that they're not warranted, hence not self-warranted.

Therefore, we can't say that what explains the privilege of privileged access is that beliefs about our own mental states are always self-warranted. Inferential beliefs about our own mental states are not self-warranted,

and some of our non-inferential beliefs are not warranted at all.

### 3c. Privileged access and consciousness

I think that what went wrong with Alston's analysis was not that it didn't apply to privileged access. We can stipulate that the privilege of privileged access consists in self-warranted belief. The problem was that Alston claimed that all our beliefs about our own mental states were self-warranted, and it's clear that quite a few aren't. So we might think that privileged access must be self-warranted belief about one's own mental states, and dismiss the anomalous phenomena as irrelevant--"After all," we might say, "these are states to which we obviously don't have privileged access."

But this conclusion is too quick. It's one thing to stipulate that privileged access is a species of self-warranted belief; it's another thing to say which beliefs about our own mental states qualify as self-warranted. The anomalous phenomena cause trouble precisely because they plant a small doubt in our minds that any of our beliefs about our mental states are self-warranted. What good is

the notion of self-warranted higher-order belief if it doesn't apply to anything? If we can't rule out these anomalous cases in a principled way, it could be that defining privileged access in terms of self-warrant would define it out of existence.

I'd be prepared to accept that self-warrant simply doesn't capture the phenomena of privileged access, but for one thing: there is a very strong intuition that privileged access, whatever it is, is a species of self-warranted belief. For example, it very strongly does seem that when I believe I'm thinking that the Mets will have a good season, I'm automatically warranted in believing it. I can't imagine any evidence which would undermine my warrant. My believing it seems to me virtually conclusive evidence that it's so. And examples of this phenomenon are legion. Denying that privileged access is a species of self-warranted belief, then, doesn't seem to be a serious option. What we ought to look for, then, is a principled way of ruling out the anomalous phenomena which falsified Alston's too-general claim. We need to find a way to make Alston correct in spirit, if not in letter.

With that in mind, let's take another look at the kinds of cases which show that not all our beliefs about our own mental states are self-warranted. First, there are inferential beliefs about non-conscious mental states; second there are false or unwarranted beliefs about conscious mental states. And it should be easy to see that the trouble both of these kinds of cases made for self-warrant has to do with the way in which the higher-order beliefs are formed.

If privileged access is non-inferential higher-order belief, then the first kind of case doesn't apply, because those beliefs aren't non-inferential. Still, the second kind of case might well apply: my belief that I am having sensations of extreme heat might be non-inferential--it might just suddenly "pop" into my mind--yet still be unwarranted. But here we can make a distinction which will be useful. It is plausible that I have beliefs about my own sensations in having those sensations consciously--after all, I have knowledge of my own sensations by virtue of being conscious of them--so we can talk of a belief we can have about a mental state by virtue of which we are

conscious of it. This is distinct from a belief we might have about a mental state which, regardless of whether or not it's inferential, is one we can't have unless we are already conscious of it. It's because I am aware of my own sensation, and aware of it as having a certain quality, that I can come to believe, without warrant, that it is a sensation of extreme heat. So believing that my sensation is one of extreme heat, although it might be non-inferential, is nevertheless one I couldn't have unless I were already conscious of my sensation's having had a certain quality. And we can say that self-warrant applies only to those beliefs we have about our own mental states by virtue of which we're currently conscious of them as we're in them.

This second feature of privileged access rules out not only the shower case, but also the cases of the non-existent tie-bearer and the apple sauce. When I believe that I've seen such-and-such a man wearing a tie, that belief is not the belief I had by virtue of which I was conscious of my original experience, and it's a belief I couldn't have had

unless I had already been conscious of that experience, given that I was at some time conscious of having it.

Thus, it seems that we can rule out the exceptions to beliefs about our own mental states being self-warranted by narrowing the range of mental states to those conscious occurrent states we're in about which we have non-inferential beliefs by virtue of which we are conscious of those states.

Even so, there are some complications for this view, two of which may be addressed quickly, but one which calls for slightly more extended discussion.

First, one might reasonably hold that creatures can be conscious of their own mental states without having any beliefs about them--that they possess some lower level of awareness which allows them to have conscious states without beliefs. Still, I am prepared to count these lower levels of awareness as beliefs, even if they may not be as conceptually sophisticated as ordinary human beliefs. Call any state which is either a belief or one of these lower-level states a schmelief. Then self-warrant holds for non-

inferential schmeliefs about our own mental states by virtue of which we're conscious of those states.[8]

One might also hold, reasonably, that being conscious of a mental state is simply the same thing as being in a conscious state, and further that there is no separate state of being conscious of one's conscious states. (See Shoemaker 1990.) I discuss this issue in slightly more detail in Chapter VII, but I'll state quickly here that this seems to be no problem either. If this view is correct, we can explain away the counterexamples in the following way: the reason self-warranted beliefs about our own mental states involve non-inferential beliefs by virtue of which we're conscious of those states is that being in those states is (in part) having those beliefs.

Finally, Dennett (1991) argues that in many cases there's simply no fact of the matter about whether a we have belief by virtue of which we're conscious of a mental state is true. If Dennett's right, it's arguable that in these

---

[8] It may be that these lower-level states can't but be non-inferential, because the creatures in them lack the requisite sophistication to form conscious inferences involving them. If so, such creatures are immune to the counterexamples I raise here.

cases there's no fact of the matter about whether we have a self-warranted belief which otherwise meets the criterion for being a case of privileged access.

Here's one of Dennett's examples: when two lights an appropriate distance apart flash at small intervals, it can seem to observers not that there are two flashing lights, but a single dot of light bouncing back and forth. One hypothesis explaining this is that the visual system at some non-conscious level represents the flashing dots as a single moving dot, and we become conscious of that representation; another hypothesis, however, explains this phenomenon in terms of the visual system correctly representing the flashing dots, and our misrepresenting this correct representation as that of a single moving dot. In the former explanation, we have a true belief about a misperception; in the latter, we have a false belief about a correct perception. What complicates the difference between the two explanations is that the time interval between the information about the flashing lights reaching the visual cortex and our conscious experience as of a single moving

dot is so small that it's hard (if not impossible) to draw a principled distinction.

I'm not sure what to say about cases like this. On the one hand, I'm inclined to think that we do have self-warranted beliefs even here, even if there's no way to determine whether or not they're true. Warrant doesn't entail truth, so we needn't demand that all warranted beliefs be true. On the other hand, I'm inclined to think that these cases are so anomalous that even if they were cases in which we lacked self-warranted belief, it would make only a tiny bit of trouble for my characterization of privileged access. What these cases show, if they show anything, is that the domain of privileged access has vague boundaries. I'm perfectly willing to accept that privileged access is a slightly blurry phenomenon--what phenomenon isn't slightly blurry? So although there may be indeterminacies at the boundaries of privileged access, we needn't be worried by them.

### 3d. Self-warrant as non-empirical authority

Throughout this chapter, I have been emphasizing the non-empiricality of privileged access. Privileged access

seems to be knowledge which doesn't depend on empirical evidence, and in particular doesn't seem answerable to it. Why is this so?

Given that privileged access is a species of self-warranted belief, the answer is obvious. Self-warranted beliefs can't be undermined by empirical evidence. A belief is self-warranted just in case having it entails that it's warranted. But if privileged access were answerable to empirical evidence, it should be possible to have that belief and not be warranted in believing it; so privileged access's being self-warranted belief explains why empirical evidence isn't relevant to the question of whether any of our privileged beliefs about our own mental states are true.

Nevertheless, I must note a complication raised by Stephen Schiffer. It seems that I have non-empirical authority over whether or not I'm having a certain pain; but it's also arguable that I don't have non-empirical authority over whether or not I have a toothache (a certain pain), because evidence might arise which shows that I have no teeth--a circumstance surely relevant to whether or not I have a toothache. So do I really have non-empirical

authority in general about what mental states I'm in? Perhaps not.

Having raised this complication, I'll put it aside. We'll return to it in Chapter VI, though, because it suggests a view of privileged access which isn't privileged knowledge of the propositions one is thinking.

## 4. Conclusion

Let's quickly summarize. Having privileged access to one of one's own mental states consists in having a non-inferential belief about that mental state by virtue of which one is conscious of it; such beliefs are non-empirically authoritative because they are self-warranted. This much, we'll see, is enough to motivate a significant part of the argument that externalism and privileged access conflict.

We need to explain these epistemic features, of course, but this task is beyond us at the moment; I'll sketch some of the issues in Chapter VII. Before knowing why privileged access is the kind of knowledge it is, however, we must know why there seems to be a conflict between it and externalism. And if we're to know that, we need to know first what

externalism is, and why it seems such a plausible view about mental content. So to see what externalism is, and why so many people think it's the best view of mental content, let's proceed to Chapter III.

# III

# EXTERNALISM

## 1. Introduction

Externalism is different things to different people. Some regard it as a thinly veiled causal theory of reference; others regard it simply as the denial of individualism, a once-dominant view about the relationship between mind and world; still others regard it as a group of essentialist claims about natural kinds; some, no doubt, regard it as all three.

My purpose in this chapter is not to say who's right about what the reference of 'externalism' ought to be (a dull project, to me), but simply present some arguments for an externalistic thesis that makes clear why one might think it leads to the paradox which is the subject of this essay.

I discuss here four arguments for externalism. The first is the argument from Twin Earth, made famous by Putnam (1975); the second is the argument from community, due to Burge (1979); the third is the argument from naturalism,

based on naturalistic considerations about intentionality, and due to such writers as Dretske (1981), Fodor (1990), and Millikan (1984); the fourth is the argument from direct reference, which is based on arguments about the reference of singular terms in 'that' clauses advocated by those who fall within the "direct reference" camp of semanticists.

Of the four, the first two seem to me the most persuasive, while the third and fourth are far less convincing. What's more important, though, is that the more persuasive arguments are far more plausible than any arguments hitherto offered by individualists about content. That's because, as we'll see, the best arguments for externalism present intuitions whose truth doesn't seem explicable on any individualistic theory of content.

Be that as it may, none of the best arguments for externalism suggest a unique thesis of externalism. Indeed, it's not obvious that there even is a unique thesis. Still, I won't go through the literature on externalism in search of a unique externalistic thesis; instead, I'll stipulate the following characteristic thesis of externalism:

> The propositional contents of the propositional attitude states of thinkers are often partly determined by factors external to them.[1]

I'll then show how well the four arguments work in demonstrating the characteristic thesis.

Now, the characteristic thesis suggests that having thoughts with propositional content entails bearing some kind of relation to one's physical or social environment, but it's unspecific as to just which relations one must bear to count as having some thought or other. Some relations suggested in the literature are quite strong; others are quite weak. We can capture this by stating that if the characteristic thesis is true, then necessarily:

If x thinks that ... y ..., then R(x,y),

where R is a suitably causal and/or socially mediated relation between a thinker x and an object y. Just how strong a thesis externalism is will depend, then, on what sort of relation R is; and various candidates for R-hood can be defended by appeal to various intuitions.

Here's how I'll proceed. After a brief prelude about concepts, I'll discuss each of the five arguments for

---

[1] I owe this way of stating externalism to lectures of Stephen Schiffer's.

externalism in turn. In the process, I'll examine individualistic alternatives and find them wanting. I'll then discuss different ways of understanding the relation R, and how it corresponds to various strengths of externalism.

## 2. A terminological prelude

I'll talk a lot about concepts in this essay; but there is as yet no consensus about what a concept is or what it is to have one. I have little of substance to say about concepts, but what I mean by the term is roughly captured by the following remarks.

My believing that snow is white depends on my having the concepts snow and white: if I lacked either of these concepts, I wouldn't be able to believe that. Also, my belief that snow is white shares content with my belief that snow is frozen water, and that thing by virtue of which they share content is my concept snow. So let's call a concept a propositional attitude "constituent" in just this sense: having a concept c is necessary for believing that ... c ..., and believing that F(... c ...) and believing that G(... c ...) share content with respect to c.

If we like, we might also posit an identity condition for concepts: a thinker's concepts c and d are identical just in case there is no extensional context F such that the thinker both believes F(... c ...) and fails to believe F(... d ...). But nothing in this essay will depend on this (plausible) identity condition being true generally.

## 3. The argument from Twin Earth

### 3a. The argument

The argument I'll offer here is adapted from Putnam's famous (1975) argument. But it's worth briefly noting the differences between the argument I offer here and Putnam's actual argument.

Putnam's argument is explicitly about linguistic meaning, and by itself entails nothing about the content of propositional-attitude states. Indeed, Putnam makes a very un-externalistic assumption about mental states with the aim of showing that linguistic meaning depends on factors external to speakers: that being in a mental state presupposes the existence of no individual other than the

subject in that state.[2] But Tyler Burge (1982) has shown how to extend the argument in a natural way to cover mental content, and I will follow his lead.[3]

Let's assume that there's a planet which is qualitatively identical to Earth--call it 'Twin Earth'--but for a single difference: while on Earth there are rabbits, on Twin Earth there are creatures which are superficially indistinguishable from rabbits, but nevertheless not rabbits.[4] (They can't interbreed with rabbits, for example.) Let's call these creatures 'twin rabbits', or simply 'twabbits'. Now, consider some Earthling we'll call

---

[2] This is Putnam's methodological solipsism (1975, 220). Also note that Putnam himself finds the assumption implausible, and that he merely uses it for the purpose of demonstrating externalism about linguistic meaning.

[3] A small point: Burge remarks (1982, 102) that the externalism he defends is 'incompatible with some of what [Putnam] says'--in particular, Putnam's statement that Earthlings and their twins are 'exact duplicates in ... feelings, thoughts, interior monologue, etc.' (1975, 224). This is misleading, and can make it seem that Putnam's statements were in error, because he did not see how the argument naturally extended to mental content. If we keep in mind that Putnam was assuming methodological solipsism merely for the purposes of discussion, we have no reason to attribute such an error to him.

[4] This way of putting the argument, save the choice of example, is based on lectures of Schiffer's.

Fred. Fred, like most of us, doesn't know anything about rabbits except what he's seen of them in zoos, read of them in books and newspapers ("Michael Jackson builds a rabbit farm in Neverland!" etc.), and so on. More precisely, Fred doesn't know the exact individuating properties of rabbits. Even so, Fred, a well educated speaker of English, is as competent as we could expect anyone to be in the use of the term 'rabbit' in English, and as capable of distinguishing rabbits from non-rabbits as any of his fellows.

Now, we're very strongly inclined to say of Fred that he has beliefs about rabbits--indeed, so strongly inclined as to think it obvious. What's more, we're not at all inclined to say that Fred has beliefs about twabbits. Fred has never come across a twabbit (there aren't any on Earth for him to come across), he's never heard of any such creature, and he doesn't know enough biology to imagine that there might be a creature which has the individuating properties of a twabbit. (This is to be expected: he doesn't even know the individuating properties of rabbits, and he knows about them.) If he visited Twin Earth, and

returned with tales of rabbits roaming the fields there, he'd be mistaken, for there are no rabbits on Twin Earth.

Because Earth and Twin Earth are qualitatively identical but for rabbits and twabbits, Fred has a twin on Twin Earth, whom we'll call 'Twfred'. Twfred is molecule-for-molecule identical to Fred, and Twfred's doings match Fred's in virtually every respect. They differ, of course, where rabbits and twabbits are concerned: as Fred reads a newspaper which says, "Michael Jackson builds a rabbit farm in Neverland!" and whose story is accompanied by a photograph of some rabbits, Twfred reads a type-identical newspaper whose story is accompanied by a photo not of rabbits, but of twabbits. Twfred is as competent in the use of 'rabbit' as his twin, and can distinguish twabbits from non-twabbits precisely as well as Fred can distinguish rabbits from non-rabbits.

Now, we're very strongly inclined to say of Twfred that he has beliefs about twabbits--indeed, so strongly inclined as to think it obvious. What's more, we're not at all inclined to say that Twfred has beliefs about rabbits. Twfred has never come across a rabbit (there aren't any on

Twin Earth for him to come across), he's never heard of any such creature, and he doesn't know enough biology to imagine that there might be a creature which has the individuating properties of a rabbit. (This is to be expected: he doesn't even know the individuating properties of twabbits, and he knows about them.) If he visited Earth, and returned with tales of twabbits roaming the fields there, he'd be mistaken, for there are no twabbits on Earth.

Of course, the above paragraph about Twfred differs from the corresponding paragraph about Fred in only two respects: I've replaced all the names for Earthly things with names for Twin-Earthly things, and conversely. This should indicate that <u>whatever</u> inclines us to ascribe rabbit beliefs to Fred inclines us to ascribe twabbit beliefs to Twfred; likewise, whatever inclines us to refrain from ascribing twabbit beliefs to Fred inclines us to refrain from ascribing rabbit beliefs to Twfred. But the only relevant difference between Earth and Twin Earth is that Earth has rabbits where Twin Earth has twabbits. On the (extremely plausible) assumption that there are facts by virtue of which Fred and Twfred believe what they actually

believe, one of these facts must be the presence of rabbits (twabbits) in Fred's (Twfred's) environment--or perhaps there not being twabbits instead of rabbits on Earth (in Twfred's case, there not being rabbits instead of twabbits on Twin Earth). Therefore, the contents of at least some of Fred's and Twfred's propositional attitude states are partly determined by factors external to them.

It should be easy to see that lots and lots of Fred's and Twfred's propositional attitude states are like this. We can run like thought experiments for 'cat', 'water', 'gold', 'tiger', 'lemon', and many other terms. Terms which denote natural kinds seem especially susceptible to this kind of thought experiment. Because Fred and Twfred were, of course, chosen solely for the purposes of illustration, the argument from Twin Earth seems to very strongly support the characteristic thesis of externalism.

### 3b. A description-theoretic objection

The argument from Twin Earth does strongly support the characteristic thesis of externalism, but it doesn't strictly entail it, and it's worth seeing why.

Individualism has it that external factors play no role in determining mental content. But the argument from Twin Earth already eliminates most versions of individualism. Individualism standardly comes in one of two flavors: there is description-theoretic individualism, according to which the content of one's concept rabbit is some (complex) property which all and only rabbits have;[5] and there is inferential-role individualism, according to which the content of one's concept rabbit is the inferential role of one's rabbit beliefs.[6]

It should be clear that inferential-role individualism can't explain the intuitions behind the Twin Earth thought experiment, because the inferential roles of Fred's and Twfred's belief states are identical. And the usual way of associating descriptions with terms in a description-

---

[5] The idea is that in thinking such thoughts as that rabbits are cute, what we are thinking is really something like the Fs are cute, where F is a property had by all and only rabbits.

[6] In this context it's useful to think of beliefs as sentences of English. Then the content of one's concept rabbit just is the inferential role of 'rabbit'--that is, the role 'rabbit' sentences play in one's inferences. For a useful discussion of inferential-role semantics, see Hartry Field (1977).

theoretic semantics won't do, either, although it will take a moment to explain why.

A description theory of content says that the propositional content of a sentence containing a denoting term t is the same as that of the sentence resulting from replacing t with a definite description the F. Thus, the propositional content of 'Rabbits are cute' is identical to the propositional content of some sentence such as 'The small, furry, pointy-eared (etc.) mammals are cute.' On the assumption that the content of rabbit is some uniqueness property, Fred and Twfred will have the same uniqueness property in mind, yet differ in their beliefs. (Unless the property itself is subject to a Twin Earth thought experiment, which would invalidate the example.) So it would seem that individualistic semantics can't account for the Twin Earth intuitions.

But this is too quick, as some individualists have pointed out. Putnam (1975) urged that the reason the Twin Earth thought experiments worked the way they did was that natural kind terms were indexical: their reference depends on the essential nature of their referents, and the thought

experiments work because the sorts of contextual factors accessible to ordinary speakers, in conjunction with environment factors, determines the reference of 'rabbit' in a world. Just as we all presumably know that 'I' refers to the speaker, we all presumably know that 'rabbit' refers to creatures in our world which meet the following description: the cute, furry creatures with pointy ears (etc.). Which creatures these are depends on the world we call home, in just the way that which person is the referent of 'I' depends on who's speaking.[7] We know that such a description doesn't determine the reference of 'rabbit'--the thought experiments show that twabbits satisfy this description--but we also know that 'rabbit' in our language refers to whatever actual creatures there are in our environment which meet that description.

---

[7] Burge (1982) has argued that natural-kind terms such as 'rabbit' aren't really indexically related to the environment: if they were, for example, we'd expect that the reference of 'rabbit' to shift with shifts of speakers,, which it doesn't. But we can suppose that 'rabbit' has indexical elements, to be revealed below, even if we accept that it's not a true indexical. (That is, nothing rides on whether 'rabbit' deserves to be classified among the indexical expressions of English.)

For an argument that natural-kind terms are true indexicals, see Joseph Almog (1981).

Given this, an individualist might take advantage of the alleged indexicality of 'rabbit' and hypothesize a description which explicitly involves an indexical element:

> 'rabbit' means 'the cute, furry, pointy-eared etc. creature in my environment'.

In an analogous way, this individualist might claim that having a belief involving the concept rabbit requires having the concept the cute, furry, pointy-eared etc. creature in my environment. Something more is required for having the concept rabbit--one must be in an environment containing rabbits--but we could then stipulate that having the concept rabbit is nothing more than being in an environment containing rabbits while having the required indexical-descriptive concept. Thus, believing that rabbits roam the fields of Earth amounts to something like the following:

> Fred believes that rabbits roam the fields iff it's the case both that Fred believes that some of the cute, furry, pointy-eared etc. creatures in his native environment roam the fields and that the cute, furry, pointy-eared etc. creatures in his environment are rabbits.

Now, not everyone need associate the same description with the term 'rabbit': there might be many different descriptions such that different people associate different

descriptions with 'rabbit'. But we can say that believing that ... rabbits ... consists in there being some indexical-descriptive concept the F such that one believes ... the F ... and that rabbits alone satisfy the F. Therefore, the individualist can claim that having rabbit beliefs are not basic: having rabbit beliefs requires having beliefs involving indexical-descriptive concepts which apply to rabbits alone. Call the propositional contents of beliefs which involve no non-basic concepts basic contents.

Let's call this theory the indexical-description theory (IDT).[8] The externalistic conclusion, strictly, does follow from IDT, but its force is so considerably weakened that we need to modify it, and for the following reason. According to IDT, having the concept rabbit involves two components: having a basic concept--some indexical description denoting rabbits in the thinker's environment--and being in the relevant environment. Thus, Fred and Twfred share all beliefs which have only basic propositional content. Of course, because the contents of the indexicals are

---

[8] IDT was so dubbed by Schiffer (lectures), and my discussion here follows his. This theory has antecedents in John Searle (1983) and Michael McKinsey (1987).

different, in a picayune sense these beliefs differ in propositional content--but they differ only in the sense that Fred and Twfred have different beliefs when each believes that he himself is hungry, and such beliefs don't have their contents determined by factors external to either Fred or Twfred.

If IDT is true, then, the basic propositional contents of belief states aren't partly determined by factors external to thinkers, and insofar as externalism says that the contents of our mental states are typically so determined, we have retained the spirit of individualism without denying the force of the Twin Earth thought experiments.

This means that if externalism is inconsistent with any form of individualism, we shall have to restate it as follows:

> Even the basic propositional contents of the propositional attitude states of thinkers are often partly determined by factors external to them.

Clearly, we must also reconstrue the argument from Twin Earth as attempting to show that even the basic propositional contents of Fred and Twfred are often partly

determined by factors external to them. The question we must ask, then, is: Should we believe IDT, or should we believe externalism (as reconstrued)?

### 3c. Externalism and IDT

The question of whether we should believe externalism or IDT is really the question of whether it's plausible to ascribe beliefs involving indexical-descriptive concepts to thinkers who have such concepts as rabbit, water, gold, and the like, solely because they have these concepts. One thing we must require of basic propositional contents, of course, is that the indexical-descriptive concepts contained therein don't themselves involve constituent concepts which are susceptible to Twin-Earth-style thought experiments. It's not obvious that this is so, but then again, it's not obvious that it isn't; so for the sake of argument let's assume so. But why should we believe that people must have such beliefs in order to count as having beliefs involving non-basic concepts?

The answer can't be, "Because otherwise the Twin Earth thought experiments would demonstrate externalism," because

that's what at issue. We need some independent reason for thinking that this is what people must believe if they are to have non-basic beliefs. And it's hard to see what such a reason could be.

One could say, "You simply wouldn't count as having beliefs about rabbits unless you also had some belief involving an indexical-descriptive concept which denotes all and only rabbits." One could hold this view, and there is no obvious way of showing it's wrong. But it's worth noting that this view plausibly commits one to denying that in communicating beliefs one is communicating their contents, because it's extremely implausible to hold that when a speaker tells a hearer (say), "Rabbits are cute," the hearer comes to know what indexical description the speaker associates with 'rabbit'. The hearer may associate her own indexical description with 'rabbit', but there seem to be no contextual factors which would help her come to believe the proposition the speaker is expressing in uttering the sentence. Given this, it would seem to be sheer luck that anyone comes to believe precisely what others do about rabbits, water, and the like.

Of course, I've ignored the possibility that an indexical description is metalinguistic, as in,

> 'rabbit' means 'the cute, furry, pointy-eared (etc.) creatures which we call "rabbit" in our native environment',

but this is a rather sophisticated concept to attribute to those who talk about rabbits. Is it required for a speaker to have a concept of rabbits that she have a concept of an environment? Is it required for a speaker to have a concept of rabbits that she know that rabbits are called anything? These are dubious requirements, perhaps too sophisticated to be plausible.

So it seems to me that IDT is a rather implausible theory on its own. If true, it could explain the Twin Earth intuitions; but we have some reasons for thinking it's not.

If these reasons are conclusive, we have no reason to doubt that the argument from Twin Earth establishes externalism, even as reconstrued.

## 4. The argument from community

### 4a. The argument

The argument I'll offer here is based on that of Burge (1979), and subsequent papers. Burge offers an argument for the communal determination of propositional content by appeal to three exhaustive cases, each of which refer to linguistic factors. The cases are as follows: (i) a speaker's partial understanding of a term; (ii) a speaker's mistaken understanding of a term; (iii) a speaker's full and correct understanding of a term. Burge argues that each of these are cases in which communal factors partially determine the contents of a speaker's attitudes.[9]

By appealing to three collectively exhaustive cases, Burge hopes to show that externalism is true for all concepts, not merely those of natural kinds. It's worth noting that his argument doesn't strictly establish that conclusion--the particular cases he cites may not generalize--but it strongly suggests it. In effect, Burge

---

[9] The reference to speakers is inessential. Insofar as we can attribute beliefs to non-speakers, analogous arguments will apply. (See note 13 below.)

offers a recipe for showing the dependence of propositional content on communal factors, and his argument depends on how well the recipe generalizes. Still, we have reason to suspect that his recipe generalizes quite well indeed.

Burge's strategy is to hold a speaker's beliefs, verbal dispositions, etc. constant, while imagining a surrounding community whose usage is such that some of the sentences she's disposed to assert have truth values different from the ones they have in the larger community. In such a situation, we'd be inclined to say not that she's gained or lost any linguistic competence, but rather that some of her beliefs are different. Let's see how this strategy works.

(i) A speaker's partial understanding of a term

In this case, a speaker knows many but not all of the correct applications of a term; she's unsure or agnostic about the rest. Let's return to Fred, who has beliefs not only about rabbits, but also about bodily states. One such state he might have is that of being arthritic, and he has many correct beliefs about arthritis: he knows it is an ailment which he can have in his joints, that it can be quite painful, that it restricts movements involving the

joints it affects, and so on. As it happens, though, he is unsure whether it applies to bones generally: he doesn't know, and hasn't considered, for example, whether arthritis is an ailment that one can have in one's thigh. In fact, he can't--arthritis afflicts joints only--but, knowing that he's generally unacquainted with the science of medicine, he doesn't feel himself to be in a position to have an opinion on the matter, and indeed has no opinion.

The correct usage of 'arthritis' is less conservative than Fred's, of course; but it needn't be so. For if the usage in Fred's community had included far more rheumatoid ailments, he would be correct in ascribing the term 'arthritis' to his thigh.[10]

Clearly, this case doesn't demonstrate that it's contingent that arthritis is a joint ailment: by definition, it's an inflammation of the joints. But then Fred couldn't be correct in believing that he has arthritis in his thigh. The only correct conclusion, then, is that he would correctly believe that some other affliction--call it

---

[10] For example, many doctors, scientists, and educated laypersons would know that the sentence 'Arthritis can be had in one's thigh' is true.

'tharthritis' can be had in his thigh. Tharthritis is what Fred's community means by 'arthritis' in the counterfactual situation. What Fred would correctly believe is that he has tharthritis in his thigh. In such a situation, of course, his use of 'arthritis' would refer to tharthritis and not arthritis. Thus, he wouldn't have any beliefs about arthritis--all his 'arthritis' beliefs would involve the concept tharthritis.

The only relevant difference between Fred's believing that arthritis can be very painful and his believing that tharthritis can be very painful is the correct range of usage of 'arthritis' in his community. Therefore, this factor partially determines Fred's 'arthritis' beliefs; as it is a factor external to Fred, this case supports externalism.

(ii) A speaker's misunderstanding of a term

Let's suppose, now, that Fred has gone over the line from agnosticism to full belief. Fred gets a pain in his thigh, and comes to believe that the pain is caused by his having arthritis there. Clearly, Fred has a misconception of arthritis, but there is no good reason to doubt that he

thereby lacks beliefs about arthritis altogether.[11] Fred has simply acquired a new belief; if he counted as having beliefs about arthritis before, he should be so counted now.

In our counterfactual situation, of course, Fred's misconception vanishes. In the counterfactual community, 'arthritis' refers to an ailment that can be had in one's thigh, so when Fred affirms 'I have arthritis in my thigh,' he affirms something that may well be true, as opposed to something that's false by definition. By the same reasoning as before, it follows that Fred does not have a false belief about arthritis in the counterfactual community, but instead a true belief about tharthritis. But Fred has in no way been altered; so the content of Fred's belief state depends on features of the 'arthritis' practices of his community.

(iii) A speaker's correct understanding of a term

Finally, let's suppose that Fred actively doubts that arthritis can be had somewhere other than in joints. Fred, of course, is absolutely correct to do so: arthritis can't be had anywhere but in joints. But in our counterfactual

---

[11] Reinaldo Elugardo (1993), though, offers a reason, and we'll discuss his argument shortly.

situation, of course, his active doubt is incorrect of tharthritis: doctors will (let's hope politely) correct him. His active doubt in the counterfactual community will be doubt that he can have tharthritis somewhere other than in joints, and this doubt will be mistaken. Once again, because Fred is in no way altered, the content of his belief state depends on the 'arthritis' practices of his community.

(iv) The conclusion

Fred's concepts can either be incomplete, mistaken, or correct; Fred has no other such concept. 'Arthritis' and Fred's concept arthritis were merely chosen for the sake of example; we could just as easily have used just about any other.[12] Burge provides a recipe for creating an externalism-entailing thought experiment: find a contingent feature of a term's application and imagine it necessary, or a necessary feature of a term's application and imagine it contingent. It seems easy to use this recipe to construct thought experiments. Therefore, if Burge's sub-arguments

---

[12] Burge (1979) briefly discusses several others: 'sofa', 'brisket', 'contract', 'clavichord', 'red'.

are each sound, then just about every concept depends for its content on communal factors.

### 4b. Some objections, with replies

I now turn to some objections to Burge's master argument. I think all of them can be met, as we'll see. Burge himself considers only one of these objections, discussing it along with a whole host of others in his (1979); I won't go into the remaining objections Burge himself raises, because they seem to me far too implausible to be worth discussing here.

#### The IDT objection

According to this objection, Burge's argument neglects to consider the possibility that Fred has an indexical-descriptive concept by means of which we are entitled to ascribe arthritis beliefs to him in one situation, and tharthritis beliefs to him in another. There will surely be some indexical description which is common to both arthritis and tharthritis, such that if Fred is an arthritis-referring community he has the concept arthritis, and has the concept

tharthritis if he is resides in a tharthritis-referring community.

Of course, this objection won't do. First of all, IDT, as we've seen, is an implausible theory. Secondly, just about the only indexical description which could plausibly have the same extension as 'arthritis' as it is used in each community is 'disease called "arthritis" by the experts'; and it surely seems possible for someone to have the concept arthritis without knowing what the disease is called, or even that it has a name. Such metalinguistic requirements seem far too sophisticated to hold for all speakers as a condition for concept possession, as we saw in our earlier discussion of IDT.[13]

---

[13] It's easy to extend the Burgean argument to non-linguistic creatures. We might characterize a dog's mental state by saying, "She thinks there's meat in her bowl," and thereby ascribe to her the belief that there's meat in her bowl. Burge could hypothesize a counterfactual linguistic community in which 'meat' refers only to red meat. Call their corresponding concept schmeat. In such a community, that dog wouldn't have the belief that there's meat in bowl, but instead the belief that there's schmeat in her bowl. Note that the dog has no beliefs about language.

**The transparency objection**

According to this objection, the occurrence of 'arthritis' in ascriptions of belief to Fred are all in transparent position; that is, they are all *de re* (as opposed to *de dicto*) ascriptions. And it's well-known that *de re* ascriptions don't specify a believer's concepts: Fred's believing *of arthritis* that he has it in his thigh plainly entails nothing about the content of his 'arthritis' concept. Of course, externalism follows trivially from *de re* ascriptions--it's by definition that they are dependent on external factors--so such ascriptions clearly have no significant consequences for content determination.[14]

---

[14] Kent Bach (1988) takes a view like this one. He argues, in effect, that when someone accepts a sentence involving 'arthritis' which denies a feature definitionally true of arthritis, viz. being a joint ailment only, that person has only a *de re* arthritis belief.

Unfortunately, Bach is evasive as to precisely why this must be so. He cites Burge's 'sofa' example, which tries to show that someone can have beliefs about sofas even though she thinks them religious artifacts not made for sitting. But Bach's claim seems to be that someone who denies that sofas are made for sitting doesn't really believe that there are sofas *as others understand 'sofa'*. I doubt this, but even if it were true, it would have no relevance to Burge's 'arthritis' example. (Surely Fred believes that other people might have arthritis as they understand 'arthritis'.)

This reply would work if it were true that all such ascriptions are de re. But why should we think they are? Burge deals with this objection ably:

> The subject's belief that he has arthritis in the thigh might be interpreted as a belief of the non-arthritic rheumatoid ailment that it is in the thigh. But it hardly accounts for the relevant attributions.... The subject thinks of the disease in a certain way. He thinks of each disease that it is arthritis. Other terms for arthritis (or for the actual trouble in his thigh) may not enable us to describe his attitude content nearly as well.[15]

Burge is pointing out that the occurrence of 'arthritis' in such ascriptions as 'Fred believes that he has arthritis in his thigh' is opaque. If we substitute some other expression co-extensive with 'arthritis' in this ascription, we may not yield a truth: for example, it may not be true that Fred believes that the rheumatoid ailment afflicting joints is in his thigh, even if that ailment just is arthritis. Even if we substitute an expression co-extensive with the actual ailment in his thigh, we may not yield a truth. This is virtually conclusive evidence that the occurrence of 'arthritis' in the ascription, 'Fred believes

---

[15] Burge (1979), 548.

he has arthritis in his thigh' is in fact opaque. So the transparency objection fails as well.[16]

## The idiosyncratic conception objection

In the passage quoted above, Burge's remarks suggest that a reason for thinking that belief ascriptions about arthritis to Fred involve opaque occurrences of 'arthritis' is that Fred thinks of the disease in his thigh and of arthritis in a particular way. But what way is that? Reinaldo Elugardo argues that, in the case of misconception, it must be some idiosyncratic description Fred associates with 'arthritis'. Thus, Fred's acceptance of the sentence 'Arthritis can be had in one's thigh' suggests that the appropriate description for Fred is 'the kind of disease one can have in one's thigh and is arthritis'. The problem is

---

[16] Stephen Schiffer points out that the hidden-indexical theory of belief reports complicates this reply. We might hold that the occurrence of 'arthritis' is transparent even though substitutivity fails because of contextual factors. Even so, the hidden-indexical theory is a version of direct reference, which in itself is thoroughly externalistic. For the transparency objection to hold, there must be a viable distinction between transparent and opaque occurrences of expressions, so that there is a real difference between believing that arthritis is in one's thigh and believing of arthritis that it is in one's thigh. The hidden-indexical theory seems to lack the resources for such a distinction.

that this description doesn't denote arthritis, so Fred's idiosyncratic conception isn't a conception of arthritis.[17]

Burge has three replies available to him here. First, Elugardo doesn't go so far as to say what Fred's idiosyncratic concept is; all Elugardo says is that the definite description 'the kind of disease one can have in one's thigh and is arthritis' somehow captures Fred's concept. If this concept just is the kind of disease one can have in one's thigh and is arthritis, it's hard to see why Fred wouldn't have the concept the kind of disease one can have in one's thigh and is tharthritis in the counterfactual situation. But if it's the kind of disease one can have in one's thigh and is called 'arthritis', Elugardo is either advocating or attributing to Burge an extremely implausible description theory, the indexical version of which was criticized in the discussion of IDT.

Second, Burge is in no way committed to claiming that in having a particular way of thinking about arthritis, Fred must have in mind a particular description of arthritis. All that follows from Fred's having a particular way of

---

[17] See Elugardo (1993). Elugardo uses Burge's 'sofa' example; I've tailored his comments to fit 'arthritis'.

thinking about arthritis is his having a disposition to think some thoughts about arthritis and not others. Having such a disposition is as much a way of thinking about arthritis as is having in mind a particular description whenever one thinks 'arthritis' thoughts. Only if we assume that the only way Fred can have a way of thinking about arthritis is by having in mind such a description as Elugardo imagines does Burge's argument fail; but this assumption is false.

Third, when Elugardo claims that Fred's conception of arthritis is captured by the description 'the kind of disease one can have in one's thigh and is arthritis', he's claiming that it's not really true that Fred believes that he can have arthritis in his thigh, because this description doesn't apply to arthritis. But because Fred sincerely affirms the sentence 'Arthritis can be had in one's thigh,' is otherwise competent in the use of the term 'arthritis', and so on, we have no pretheoretic reason at all to deny that he believes (falsely) that he can have arthritis in his thigh. Because Fred wouldn't have that belief if he had the

idiosyncratic conception in mind, it's hard to believe he actually has that conception in mind.

Therefore, the idiosyncratic conception objection to externalism fails as well.

### 4c. Conclusion

Burge's argument for externalism differs from Putnam's with respect to the type of externalistic relation one must bear to the environment in order to have the thoughts one actually has: Burge's relation is social, while Putnam's is environmental. Burge's argument also seems to apply with much greater generality--Putnam's doesn't obviously extend beyond natural kind terms, and probably doesn't cover all such terms. (Consider 'environment', for example: how do we run a Twin argument for that term?) But both arguments seem very effective in demonstrating that content is partially determined by external factors.

One final note: it's been popular to appeal to 'narrow' content, an alleged kind of content which does supervene on the intrinsic physical states of thinkers.[18] Nothing I've

---

[18] Fodor (1980) is the classic defense of narrow content. He reconsiders the view in his (1994).

argued here precludes the existence of narrow content, hence the truth of individualism with respect to some kind of mental content. But I'm not prepared to accept that propositional content, the full-blooded kind of content we've been talking about all along, is individualistic. Narrow content may or may not exist, but its existence affects nothing I've argued here.

## 5. The argument from naturalism

If we assume, plausibly, that there is some natural feature of human beings such that our having this feature explains our being in intentional states, then we can't avoid the conclusion that there is some natural fact by virtue of which our mental states are about the things they are actually about. Thus, Fred's belief that rabbits are cute is about rabbits by virtue of some sort of relation Fred bears to rabbits. Just what kind of relation could that be?

Naturalistic theorists of content offer various answers to this question: one might hold, as Fodor (1990) has, that having a belief about rabbits depends upon one's underlying belief state tokens being asymmetrically dependent upon

rabbits--that is, upon their being such that they wouldn't be caused by non-rabbits unless they were caused by rabbits. Or one might hold, as Dretske (1981) does, that having a belief about rabbits depends upon being in a state which carries information about rabbits, which is something like being a reliable indicator of rabbit-presence. Or one might hold, as Millikan (1984) does, that having a belief about rabbits depends upon the proper function of the underlying belief state--very roughly, upon its being a state which is by design supposed to indicate the presence of rabbits. All of these theorists claim that, ultimately, having the concept rabbit depends upon being causally related to rabbits in some more or less robust way.

Why is this so? Because, on these theories, being asymmetrically dependent upon rabbits, being a reliable indicator of rabbit-presence, and being a state with the proper function of indicating rabbit-presence are all themselves robust causal relations to rabbits: presumably, being asymmetrically dependent on rabbits is not being asymmetrically dependent on rabbits or any rabbity non-rabbits, and the same presumably holds, mutatis mutandis,

for all the other relations. Here, we may include social factors as being suitably (though less perspicuously) causal: I can become causally related to electrons in the relevant way by attending a physics lecture, or by reading a popular book about physics, or by listening to people who have more or less knowledge about electrons, as well as by making use of a cloud chamber. Social relations to objects are causal because they are constituted by (typically byzantine) causal links to them. (For example, my utterances of 'Abraham Lincoln' are about Abraham Lincoln at least partly because I learned the term 'Abraham Lincoln' from people who learned the term from people who learned the term from people who ... who actually knew Abraham Lincoln.) No doubt, these theories will have to delicately explain how my belief that Abraham is dead can have that content if it's having that content depends (in Fodor's case) on being asymmetrically dependent upon someone who's been dead for over a century; but there's no reason in principle to doubt that such an explanation is possible.

We might portray the general argument from naturalism as basically a "how else?" argument: how else could we have

states about things in the world unless we were somehow causally related to them? Such an argument in itself doesn't entail externalism, for externalism is a view about the determination of content, not a view about what explains content. Individualism is compatible with a causal explanation of intentionality, although it isn't compatible with such an explanation conjoined with the Putnam-Burge intuitions. (IDT notwithstanding.) Still, we needn't suppose that 'the argument from naturalism' is a singular term as much as it is an umbrella term for naturalistic views of content. As such, its plausibility derives from the plausibility of such naturalistic views as Fodor's, Dretske's, and Millikan's. And these views are all quite explicitly externalistic, insofar as having such concepts as rabbit depends upon having causally interacted with rabbits.

I'm sympathetic to the project of naturalizing content, although I have some reservations about all the current theories. My mentioning them here is intended solely to show that the Putnam-Burge intuitions, though powerful, aren't the only reasons one might have for believing externalism.

## 6. The argument from direct reference

The argument from direct reference for externalism is banally simple:

> Direct reference provides the correct semantics of belief reports. If direct reference provides the correct semantics of belief reports, then externalism is true. Therefore, externalism is true.

This argument is obviously valid. But is it sound? It's very easy to demonstrate the second premise--direct reference has particularly strong externalist consequences--so the burden of one who would claim that the argument from direct reference is sound is to demonstrate the first premise.

Direct reference is controversial (although widely accepted), and there are no decisive arguments that it is a correct account of belief reports. My purposes here are modest. I hope to show only two things: that direct reference entails externalism, and that there are some good reasons for preferring it to its major alternative, Fregeanism.

In the next section, I sketch a framework for understanding a semantic theory of belief reports, one in

which we can easily compare direct reference to Fregeanism, and explain both direct-reference and Fregean accounts of belief reports in this framework. In Section 6b, I show how direct reference entails externalism. I conclude with Section 6c, in which I offer some prima facie good reasons for preferring direct reference to Fregeanism. Again, I don't expect these reasons to strike one as decisive; they need merely make one queasy about accepting Fregean views of the semantics of belief reports.

## 6a. Two theories of belief reports

### (i) A Framework

The following three sentences are examples of belief reports:

> Eric believes that Jim Neighbors was a fine singer.
>
> Patty believes that there are denumerable models of arithmetic.
>
> Marvin believes that 'the Matterhorn' denotes a kind of goat.

It's natural to represent these sentences as relations between believers and things believed, because of natural inferences we can draw from them:

> Eric believes that Jim Neighbors was a fine singer, and so does Patty. So there's something that both Eric and Patty believe.
>
> Marvin believes that 'the Matterhorn' denotes a kind of goat. What Marvin believes is false. Therefore, that 'the Matterhorn' denotes a kind of goat is false.

Each of these inferences seem to depend for their validity on 'that' clauses being referential singular terms. So, following Schiffer (1987), I'll take belief reports to be sentences of the form

> x believes y,

where 'x' ranges over believers and 'y' ranges over things believed. This treats belief reports as expressing relations between believers and the referents of 'that' clauses: believing is just that dyadic relation. We can then stipulate that a proposition just is the referent of a 'that' clause. To illustrate, let Patty be the value of 'x', and let 'B' denote the belief relation. Then the following is a perspicuous rendering of the logical form of the second belief report mentioned above:

> B(Patty,the proposition that there are denumerable models of arithmetic).

It's worth noting, as Schiffer does,[19] that some theories of belief reports will deny this assumption: Davidson's paratactic theory of belief reports represents the logical form of 'Fred believes that that rabbit is cute' as

Fred believes that.That rabbit is cute,

where 'that' is a demonstrative referring to an utterance of the sentence following the dot; also, some Fregean theories of belief reports have it that the logical form of that sentence is

∃m(m is a mode of presentation of that rabbit &
B(Fred,the proposition that m is cute)).

On neither of these theories does 'that that rabbit is cute' function as a referential singular term. Happily, though, I won't be discussing the paratactic theory, and it's easy to accommodate Fregean theories which make the claim above into our framework. On such theories, the 'that' clause corresponds to a proposition which expresses the propositional content of Fred's belief, even though it doesn't, strictly speaking, refer to that proposition.

---

[19] Schiffer (1987), 9.

These theories won't have anything essentially different to say about what propositions are apart from their not being the referents of 'that' clauses; that is, they'll say pretty much what other Fregean theories say about what propositions are. Therefore, whatever we say about the nature of propositions won't beg any questions against these theories.

(ii) What's common between the two accounts

Both the direct reference and the Fregean account of belief reports will have it that the reference of a 'that' clause is determined by its syntax and the reference of its constituent terms. What's more, we can accept that the two accounts will agree as to the reference of all constituent terms in a 'that' clause besides singular terms. For simplicity, I'll allow only two kinds of terms in a 'that' clause: singular terms and predicates, and stipulate that the two accounts agree that the reference of a predicate is a property. We can then represent the proposition expressed by a sentence which contains no logical operators as an ordered pair consisting of an n-place property and an n-tuple of objects, each object corresponding to the referent of a singular term, where the order of objects is determined

by the order of occurrences of singular terms in the sentence. For example, the proposition expressed by the 'that' clause contained in the sentence 'Martha believes that Michael Jordan is balder than Scottie Pippen' can be represented by the ordered pair

> <<R('Michael Jordan'),R('Scottie Pippen')>,x's being balder than y>,

where R(x) is a function mapping singular terms onto their referents.

Note that in the above ordered pair, the reference of 'Michael Jordan' and 'Scottie Pippen hasn't been made explicit. That's because the direct reference and Fregean accounts of belief reports give very different answers to the question of the reference of singular terms in 'that' clauses. Let's see what the two accounts say about this.

(iii) The Fregean account of reference in 'that' clauses

According to the Fregean account of belief reports, the reference of singular terms in 'that' clauses are modes of presentation of what those singular terms would refer to

were they to occur in ordinary unembedded sentences.[20] We can take it that 'Michael Jordan' refers to Michael Jordan in the sentence 'Michael Jordan is balder than Scottie Pippen.' But the Fregean wouldn't say that that's the reference of 'Michael Jordan' in the sentence 'Martha believes that Michael Jordan is balder than Scottie Pippen.' That's because if it were, then any singular term which also happens to refer to Michael Jordan could be substituted for 'Michael Jordan' in that sentence without changing the proposition referred to: recall that we're assuming that the reference of a 'that' clause is determined by its syntax and the references of its constituent terms. This would mean that if

> Martha believes that Michael Jordan is balder than Scottie Pippen

is true, then

> Martha believes that Air Jordan is balder than Scottie Pippen

must also be true. But Martha might be unaware that 'Air Jordan' (colloquially) refers to Michael Jordan; indeed, she might be under the misapprehension that 'Air Jordan' refers

---

[20] The classic account, of course, is from Frege (1952).

to Michael Jordan's brother. This makes doubtful the truth of the second belief report, and intuitively has no bearing on the truth of the first; but it would seem that the two reports must both be true or false together if singular terms in 'that' clauses refer to what they refer to in ordinary contexts.

The Fregean is deeply impressed by this difficulty; so she concludes from this that 'Air Jordan' and 'Michael Jordan' don't refer to the same thing when they occur within 'that' clauses. The Fregean reasons that in having a belief about Michael Jordan, Martha has a way of thinking about Michael Jordan, and this way of thinking is, we'll say, a mode of presentation of him. The reason she believes that Michael Jordan is balder than Scottie Pippen but doesn't believe that Air Jordan is balder than Scottie Pippen is that 'Michael Jordan' and 'Air Jordan' refer to different modes of presentation of Michael Jordan in her thinking, and that only one of the two modes of presentation--the one referred to by 'Michael Jordan'--occurs in a proposition she believes. The Fregean would thus represent the semantic

content of the belief report 'Martha believes that Michael Jordan is balder than Scottie Pippen' as follows:

$B(Martha, \langle\langle m_1, m_2\rangle, x\text{'s being balder than } y\rangle$

where $m_1$ and $m_2$ are the appropriate modes of presentation of Michael Jordan and Scottie Pippen respectively.

One problem for the Fregean theory is that it often isn't at all clear just <u>what</u> mode of presentation a believer has in mind when one ascribes a belief to her; despite this, it seems possible to make correct ascriptions. One way of addressing this problem is to weaken the theory slightly: instead of saying that the content of one's report picks out definite modes of presentation, the Fregean can say that one's report says merely that <u>there are</u> modes of presentation such that the proposition the believer believes contains them.[21] According to this view, the content of 'Martha believes that Michael Jordan is balder than Scottie Pippen' is

$\exists m_1 \exists m_2 (m_1$ is a mode of presentation of Michael Jordan & $m_2$ is a mode of presentation of Scottie Pippen & $B(Martha, \langle\langle m_1, m_2\rangle, x\text{'s being balder than } y\rangle))$.

---

[21] Graeme Forbes (1987) takes this view.

This is the view which denies that 'that' clauses, strictly speaking, are singular terms. Still, it's obvious that this theory differs from the standard Fregean theory only slightly, and not at all with respect to the nature of propositions.

(iv) The direct reference account

The direct reference account differs substantially from the Fregean in what it says about the reference of singular terms in 'that' clauses. According to this account, singular terms don't refer to one thing in ordinary unembedded sentences and to another thing in 'that' clauses; instead, they refer in 'that' clauses to whatever they refer to ordinarily. So in the sentence 'Martha believes that Michael Jordan is balder than Scottie Pippen,' 'Michael Jordan' refers to Michael Jordan himself, not a mode of presentation of him. Thus, we can represent the content of the sentence as follows, according to the direct reference account:

> B(Martha,<<Michael Jordan,Scottie Pippen>,x's being balder than y>).

It's easy to see that on this account, if the sentence 'Martha believes that Michael Jordan is balder than Scottie Pippen' is true, then the sentence 'Martha believes that Air Jordan is taller than Scottie Pippen' must be true as well, because 'Air Jordan' also refers to Michael Jordan. The problem with this view is that Martha may not know that 'Air Jordan' refers to Michael Jordan, or may believe it refers to an entirely different person. This raises doubt as to whether the belief report 'Martha believes that Air Jordan is balder than Scottie Pippen' is true.

Direct reference theorists try to get around this problem in one of two ways. Some bite the bullet, accepting that if one sentence is true, so is the other; others say that belief reports express triadic relations between believers, propositions, and modes of presentation.[22]

I won't be concerned with the plausibility of such moves here; suffice it to say that they have their own merits and defects. It should be clear, however, how stark the difference is between the direct reference and Fregean

---

[22] For the bite-the-bullet move, see Nathan Salmon (1986). For the triadic-relation move, see Schiffer (1977), and Mark Crimmins and John Perry (1989).

accounts of belief reports when it comes to the reference of singular terms in 'that' clauses.

### 6b. Why direct reference entails externalism

We know that externalism says that the mental contents of propositional-attitude states are partly determined by external factors. It's easy to see that one such factor, according to direct reference, is the existence of objects referred to by singular terms contained in the 'that' clauses of true belief reports. Here's why. Assume that the sentence 'Eric believes that Jim Neighbors was a fine singer' is true; then the following relation holds if the direct reference account of belief reports is correct:

B(Eric,<<Jim Neighbors>,x's having been a fine singer>.

Now let's ask the following question: could Eric believe that proposition in a world in which Jim Neighbors doesn't exist or in a world in which 'Jim Neighbors' refers to someone other than what it refers to in the actual world? Clearly, the answer should be No. That's because if Jim Neighbors doesn't exist, then 'Jim Neighbors' doesn't refer to anything; so there isn't any proposition containing (a set containing) Jim Neighbors for Eric to believe. And if

'Jim Neighbors' refers to someone else (call this person 'Phil'), then what Eric believes wouldn't be that proposition but instead the following:

<<Phil>,x's having been a fine singer>.

In each case the 'that' clause 'that Jim Neighbors was a fine singer' wouldn't refer to the proposition it actually refers to, so Eric wouldn't believe what he actually believes were Jim Neighbors not to refer to 'Jim Neighbors'.

The existence of Jim Neighbors is obviously a factor external to Eric: if he had lived on Twin Earth instead of his twin Tweric, then he would've believed the proposition

<<Twjim Neighbors>,x's having been a fine singer>,

not the proposition he actually believes.

### 6c. Direct reference vs. Fregeanism

Why should we believe one account rather than the other? There are many, many arguments each way, but because my sole concern here is to make a <u>prima facie</u> case for direct reference, I'll mention only two arguments for it. The first, due to Kripke (1980), I'll call <u>the argument from rigidity</u>; the second, whose ultimate origin I've yet to track down, I'll call <u>the argument from anaphora</u>.

(i) The argument from rigidity

Recall that the standard Fregean account of belief reports has it that the proposition Eric believes by virtue of which 'Eric believes that Jim Neighbors was a fine singer' is true is

<<m>,x's having been a fine singer>,

where m is a mode of presentation of Jim Neighbors. But what are modes of presentation?

For the sake of discussion, let's assume that modes of presentation is that they're uniqueness properties--properties by virtue of which objects satisfy definite descriptions. (Other notions of modes of presentation exist, but Kripke's objections are easily adaptable to fit them.) This is one natural way of explaining why the Fregean account of belief reports can be understood as one facet of a general description theory of the meanings of singular terms. For any definite description The F, let THE F be the uniqueness property it denotes. Then the proposition Eric believes is

<<THE F>,x's having been a fine singer>.

Kripke (1980) offered a prima facie compelling objection to this theory. He noted that if the content of a sentence containing the name 'Gödel' was the same as that of the sentence resulting from replacing that name with some description uniquely true of Gödel, say, 'the man who proved the incompleteness of arithmetic', then if it turned out that some other person besides Gödel had actually proved the incompleteness of arithmetic, any assertion we make involving the name 'Gödel' would be about some person other than Gödel. But when we use the name 'Gödel', of course, we mean no person other than Gödel. So the Fregean theory can't be correct.[23] A sentence containing a proper name can't express the same content as the sentence resulting from replacing it with a definite description, because, as Kripke put it, names are rigid, where rigidity is understood as follows: "A designator rigidly designates a certain object if it designates that object wherever the object exists,"[24] where 'wherever' means 'in whatever possible world.'

---

[23] See Kripke (1980).

[24] Ibid., 270.

Kripke's argument wasn't made for belief contexts: his argument had to do with ordinary sentences and the question of what proper names in those sentences meant. When we apply it to belief contexts and the question of what names refer to in 'that' clauses, his argument becomes immune to a certain compelling objection. Kripke originally argued that the reason names are rigid was that modal sentences such as 'It's possible that Gödel didn't prove the incompleteness theorem' are intuitively true, but would be necessarily false if 'Gödel' were synonymous with 'the man who proved the incompleteness theorem.'

That argument isn't decisive, because it's arguable (if not obvious) that if names are synonymous with descriptions, then in modal contexts the names should not be understood as being within the scope of the modal operator. But if we move to belief contexts, this objection becomes far less plausible. We can't read names which occur in 'that' clauses as really occurring outside the scope of 'that' clauses without doing in Fregeanism: for we could stipulate that 'Eric believes that Jim Neighbors was a fine singer' is synonymous with 'Jim Neighbors is such that Eric believes

him to have been a fine singer', but that would make belief reports de re, contrary to the Fregean proposal.

Even so, we still shouldn't say that Kripke's argument as applied to belief reports is decisive. The reason is that the Fregean might well agree that if 'Jim Neighbors' referred to someone else or to no one at all, Eric's belief would be about a different person or about no one; but what's relevant is that Eric's belief would be the same--that is, he would believe the same proposition. Once we distinguish between a mode of presentation and what the mode of presentation is a mode of presentation of, we can readily accept Kripke's intuition about reference without granting that it applies to belief contexts. There just is an intuition that Eric could still believe what he actually believes even if it turns out that Jim Neighbors never existed; and the Fregean account of belief reports captures this intuition.

(ii) The anaphora argument[25]

The following sentence has two plausible readings:

---

[25] I don't know who was the first to offer this argument. Michael Dummett (1973, Chapter 9) tries (in my view, unsuccessfully) to meet it.

Fred loves Martha, and he loathes Paul.

On one reading, 'he' refers to some contextually definite person who may, after all, be Fred; on the other reading, however, 'he' can only refer to Fred. That is, it's part of the meaning of the sentence on the second reading that 'he' refers to Fred. If we read it in the second way, then we can preserve meaning by shortening it to

Fred loves Martha and loathes Paul.

If we read it in the first way, however, then the shortened sentence means something entirely different. I'll say that 'he' is anaphoric on 'Fred' if it's part of the meaning of the sentence that they co-refer.

This makes dramatic trouble for the Fregean theory. For consider the following sentence:

> Few people believed that Jim Neighbors was a fine singer, even though he really was.

The occurrence of 'he' in the above sentence can really only be read as being anaphoric on 'Jim Neighbors', which entails that they must co-refer. But the Fregean account of belief reports has it that, in that sentence, 'Jim Neighbors' refers to a uniqueness property, and the constituent sentence 'he really was' is simply elliptical for 'Jim

Neighbors was really a fine singer', which simply isn't true if 'he' refers to a uniqueness property. Indeed, the only thing that the constituent sentence can mean is that Jim Neighbors was really a fine singer, and that can only be if 'he' refers to Jim Neighbors. Therefore, if 'he' is anaphoric on 'Jim Neighbors', the above sentence can't mean what the description theorist says it means and be true (and it is true).

As far as I can tell, the only way out of this problem is to assert that the occurrence of 'Jim Neighbors' in the sentence above is transparent, and so that the beliefs are always de re. But this seems hardly plausible at all.

If this argument is correct, then the reference of names in belief contexts must be identical to their reference in other contexts, which is just to say that the truth of a belief ascription involving a proper name entails that the name refers.

Now, the argument from anaphora really does seem decisive; barring counter-intuitive treatments of anaphora, it would seem that names (and perhaps other singular terms) in 'that' clauses must refer to whatever they refer to

outside of 'that' clauses. While this argument doesn't establish direct reference--it may be false even though the argument establishes what it establishes--it does offer a *prima facie* compelling reason to prefer it to Fregeanism.

## 7. Externalism's consequences

Let's briefly summarize. In Section 1, I characterized externalism as the following thesis:

> The (basic) propositional contents of the propositional attitude states of thinkers are often partly determined by factors external to them.

In Sections 3 and 4, I appealed to arguments of Putnam and Burge to demonstrate externalism; in Section 5, I briefly sketched naturalistic considerations which support externalism; and in Section 6, I argued that semantic considerations support a particularly strong version of externalism--viz., that true belief reports involving embedded proper names entail the existence of their referents.

But what do I mean by a 'strong' version of externalism? Externalism, as characterized above, says nothing precise about the factors external to thinkers that partly determine the propositional contents of their

thoughts. What is an 'external' factor, and how does it determine propositional content?

A content-determining factor of an individual's propositional-attitude state is external if it fails to supervene on the intrinsic physical states of that individual. This means that the individual's being in its actual intrinsic physical states doesn't entail whether or not that factor obtains. For example, let's suppose that the property of having causally interacted with rabbits partly determines the contents of states involving the concept rabbit. By Putnam's argument, this property is a factor external to thinkers, because their intrinsic physical states don't entail whether or not they've causally interacted with rabbits. (Recall that Fred and his twin share all their intrinsic physical properties by virtue of being molecule-for-molecule identical.)

Given this, what sort of externalistic factors should we expect to partly determine the contents of our thoughts? This is a difficult question to answer, and I won't offer anything precise here. Instead, I'll raise considerations which suggest three different kinds of externalistic factor,

and say a little about the kinds of externalism that result. It may be that there is no single interesting externalistic generalization which holds of all propositional-attitude contents--in fact, I think this highly likely--but I'll ignore this possibility here, because my purpose isn't to defend one or another kind of externalistic factor as being most general, but merely show how they can lead to paradox.

In any event, the three considerations below support three distinct views about the role of the thinker's environment in content determination. These views come in distinct degrees of strength: one entails the other two, and another is entailed by the other two.

### 7a. Super-weak externalism

If we want to know what externalistic factors, strictly speaking, follow from the Putnam and Burge arguments, it's hard to give an answer that's terribly interesting. At best, we get the following externalistic factors:

(Putnam) Fred's thinking that ... c ... depends on there not being indistinguishable non-c's instead of c's in Fred's environment.

(Burge) Fred's thinking that ... c ... depends on Fred's not having lived in a community with linguistic practices such that the terms

> contained in true belief reports by virtue of which Fred thinks that ... c ... fail to apply to all and only c's, instead of his actual community.

For example, Fred's thinking that rabbits are cute entails that Fred hasn't lived on Twin Earth, thereby referring to and thinking about twabbits instead of rabbits. And Fred's thinking that he has arthritis in his thigh entails that he hasn't lived in a community whose use of 'arthritis' is such that Fred's attribution of the term 'arthritis' to the ailment in his thigh is true instead of false.

Thus, the externalistic relation R, according to what I'll call super-weak externalism, is a weak relation between thinkers and things. Neither the Putnamian nor the Burgean relations entail that there are rabbits or that Fred has lived in some linguistic community; all it entails is that there aren't twabbits instead of rabbits, and that Fred hasn't lived in a linguistic community such that ... instead of his actual linguistic community.

### 7b. Strong externalism

Super-weak externalism identifies the externalistic relation R with the weakest possible relations consistent

with the arguments of Putnam and Burge. But we needn't be so circumspect. It's surely plausible that there's a reason why the mental contents of one's propositional attitude states can vary while one's intrinsic physical states don't. Strong externalism offers a set of considerations which partly explain this variation.

Naturalists about content offer one or another causal account of content determination: what explains why my beliefs about water are beliefs about water is some fact about my causal interactions with water. We surely shouldn't expect that super-weak externalism holds for no reason at all, that it's just a primitive fact. We should expect there to be some explanation for why Fred's believing that rabbits are cute entails that there aren't twabbits instead of rabbits in Fred's environment, for example. And a causal theory of content determination offers just the right sort of explanation.

I've already discussed some naturalistic theories of content determination, and it's easy to see that they have the resources to explain why Fred's having the concept rabbit depends on there not being twabbits instead of

rabbits in Fred's environment. On Fodor's account, it's because Fred's tokens of 'rabbit' wouldn't be caused by twabbits unless they were caused by rabbits; on Millikan's account, it's because Fred's 'rabbit' representations have the proper function of indicating rabbits and not twabbits; on Dretske's account, it's because Fred's 'rabbit' representations carry information about rabbits but not twabbits.

This suggests that having concepts depends on bearing a suitable causal relation to things in the concept's extension; thus, if a causal theory of content determination of the sort offered by Fodor, Millikan, or Dretske is true, we should expect that the following relation to hold between Fred and rabbits, given that he thinks rabbit thoughts:

> Fred's thinking that rabbits are cute depends on Fred's having causally interacted with rabbits.

This is a particularly strong version of externalism, and a controversial one. Nevertheless, it's a version of externalism, and one that many people believe. I call it strong externalism.

Strong externalism has as many varieties as there are causal theories of content determination. But one variety

of strong externalism should be mentioned, if only because of its notoriety: it's what we might call the causal-historical theory of reference. This theory came out of two ideas in the theory of reference: (i) Kripke's (1970) idea that the reference of a name is fixed by an initial "baptism", and "knowledge" of the reference of a name was then passed on from speaker to speaker by intentions to refer to whomever some speaker referred to in using the name; (ii) the idea, due independently (?) to Putnam (1975) and Kripke (1980), that natural-kind reference is determined by the "hidden natures" of the substances to which speakers intended to refer in using those terms. The causal-historical theory of reference tends to be simply identified with externalism, because externalism is typically identified with the views of Putnam and Kripke on proper names and natural-kind terms. But we shouldn't identify the two. For one thing, it's far from clear that the causal-historical theory of reference, understood as ideas (i) and (ii), is even true;[26] for another, externalism as I've stated and defended it is independent of any particular

[26] For some compelling reasons, see Joseph LaPorte (1996).

theory of how terms get their reference, or of how the ability to make such reference is transmitted from speaker to speaker.

Nevertheless, the causal-historical theory of reference is worth noting also because of the way it connects externalism with direct reference. Strong externalism in fact entails the version of externalism which follows from a direct reference account of belief reports. If Fred's thinking that Michael Jackson built a rabbit farm depends on Fred's having causally interacted with Michael Jackson, then Fred's thinking that Michael Jackson built a rabbit farm depends on Michael Jackson's existence. (Fred's having causally interacted with Michael Jackson entails Michael Jackson's existence.) I won't treat the version of externalism which follows from direct reference separately, for it's but a slightly weaker form of strong externalism.

As we'll see in the next section, strong externalism is the strongest variety: it entails not only super-weak externalism, but also what I call weak externalism.

### 7c. Weak externalism

One might be uncomfortable with strong externalism because it follows simply from one's having beliefs involving a natural-kind or proper-name concept that that concept is instantiated; we'll see in Chapter IV a more sophisticated way of articulating this discomfort, but for the moment we need only see that one might find strong externalism uncomfortable for this reason alone. One might think that it's possible to have beliefs involving the concept rabbit, for example, even if there turn out not to be any rabbits--if, for example, one has been having vivid hallucinations as of rabbits instead of genuine perceptions.

This discomfort was addressed by Tyler Burge (1982), who argued that it in no way conflicts with the intuitions supporting externalism:

> [I]t is logically possible for an individual to have beliefs involving the concept of water (aluminum, and so on), even though there exists no water. An individual or community might (logically speaking) have been wrong in thinking that there was such a thing as water. It is epistemically possible--it might have turned out--that contrary to an individual's beliefs, water did not exist.... If by some wild communal illusion, no one had ever really seen a relevant liquid in the lakes and rivers, nor had drunk such a liquid, there might still be enough in the community's talk to distinguish the

> notion of water from that of twater and from other candidate notions. We would still have our chemical analyses, despite the illusoriness of their object. (I assume here that not all of the community's beliefs involve similar illusions.)[27]

Burge here contends that I might have the concept water even though there was no water in my environment and no one had never causally interacted with water; his reason is that there are extant chemical analyses which we falsely attribute to a non-existent liquid, and it's reasonable to think we count as having beliefs involving the concept water and not the concept twater because the following is true:

> The chemical hypotheses involving 'water' are such that they would be true iff 'water' referred to water.

But this isn't quite correct. If it were, then if there were twater in our environment instead of water and scientists had offered a chemical analysis of twater which was true of water only, it wouldn't be the case that we had the concept water instead of twater. The right thing to say in that scenario is that the chemical analysis is mistaken. But the only difference between that scenario and Burge's hallucination scenario is that in the latter there isn't any

---

[27] Burge (1982), 115-6.

liquid in our environment. And why should that difference be relevant?

What Burge needs here is a reason for thinking that the chemical analyses are true analyses of water instead of false analyses of twater. What reason could this be? As I see things, it could only be that super-weak externalism is true: it's not the case that there is twater instead of water in our environment. Burge's hypothesis is slightly stronger than super-weak externalism, yet weaker than strong externalism. It's best stated as follows:

> Fred's believing that rabbits are cute depends either on there having been rabbits in Fred's environment or on rabbithood having been hypothesized in Fred's community.

For the reasons given above, we must cash out 'rabbithood having been hypothesized in Fred's community' as

> there not having been indistinguishable non-rabbits instead of rabbits in Fred's environment and there having been a scientific practice of identifying the reference of 'rabbit' with those things having all and only the determining properties of rabbits.

This version of externalism, in my view, best captures the considerations Burge raises.

## 8. Externalism and vacuity

One interesting thing worth noting about externalism is that if a thinker is in a state such that some externalistic relation fails to hold between it and the world, and no similar relation holds in its stead, then the thinker is in a state which lacks content altogether. This is clearest in the direct reference version of externalism: if Hillary Clinton doesn't exist, then 'Hillary Clinton' is vacuous, so there is simply no belief that Hillary Clinton is from Arkansas, by virtue of there being no proposition containing Hillary Clinton. But, analogously, if one bears no externalistic relation to water or anything else like it, even though one uses the term 'water' in ways otherwise similar to ours, then one's belief states which one would express by using such sentences as 'Water is a liquid,' 'Water quenches thirst,' and so on, all lack content.

This is easy to see in the case of strong externalism and weak externalism. But super-weak externalism is an exception. If thinking that rabbits are cute depends on there not being indistinguishable non-rabbits in place of rabbits in one's environment, then there is simply no way

for one's belief state to lack content by failing to bear the appropriate externalistic relation: clearly, if one isn't related to anything in the universe, then of course there aren't indistinguishable non-rabbits in place of rabbits in one's environment. So that one fails to be related to anything doesn't falsify any super-weak externalistic consequence.

Of course, one might chafe at this conclusion. It surely ought to be possible to have vacuous belief states, one might think; but to this extent, one is claiming that super-weak externalism is too weak to be plausible. Still, I won't go into this in detail; I'll discuss some implications of super-weak externalism for the paradox in Chapter VI.

## 9. Segue

I've offered four arguments for externalism. Perhaps, in the final analysis, none of them work; but I've tried to show that two, at least--Putnam's and Burge's--are highly resistant to the standard objections. For the sake of argument, though, I'll take it that these arguments put the burden of proof on those who'd disagree with externalism.

We can take these arguments to establish externalism until an argument comes along which refutes them.

I suggested in Section 7c that strong externalism has a discomfiting consequence: that having beliefs involving a natural-kind or proper-name concept entails that the concept is instantiated. What, exactly, is so discomfiting about this consequence? In the next chapter, I discuss several arguments which try to articulate this discomfort. These arguments don't in any way refute the arguments for externalism--although some take them as reductios of the thesis of externalism. If any work, they establish that equally plausible considerations can be used show that externalism is false. If any work, then, externalism plus these plausible considerations lead to paradox.

# IV

## ARGUMENTS FOR THE INCOMPATIBILITY THESIS

In this chapter I consider arguments for the thesis that privileged access and externalism are incompatible. (Henceforth, I'll call this thesis (IT).) Such arguments are uncommon, because those who consider the incompatibility tend to assume that at least a prima facie sound argument for it can easily be made. In fact, I have found only three explicit arguments in the literature: McKinsey (1991), Brueckner (1990), and Boghossian (1989).[1]

This assumption is unwarranted, however, because finding a prima facie sound argument is far from a simple matter. In the first three sections, I discuss the three explicit arguments, and show that not one of them is plausibly sound. I then discuss, in Section 4, various ways

---

[1] Brian Loar (1992) offers an argument for (IT) which seems quite different from the three discussed here. His argument is very difficult to understand, mainly because of his somewhat cryptic presentation. But insofar as it is plausible, I think it boils down to what I call the Improved McKinseyan Argument.

in which the arguments might be improved; as we'll see, improving them so that they are plausibly sound requires some subtlety. Still, they can all be improved in such a way that they're plausibly sound; but we'll see that they all converge into one another to form a master argument for the incompatibility of externalism and privileged access.

I conclude, in Section 5, with a related issue which will arise in Chapter VI. I show that Putnam's notorious "proof" that we can't be brains in a vat follows directly from the main premises of the master argument. This will be important later on.

For now, though, let's see how the arguments work.

## 1. McKinsey's Argument

McKinsey's (1991) argument focuses on the following three claims, where Oscar is some arbitrary person ignorant of chemistry and E is an appropriate externalistic fact of the sort discussed in Chapter III, Section 7:

(A) Oscar knows a priori that he is thinking that water is wet.

(B) The proposition that Oscar is thinking that water is wet necessarily depends on E.

(C) The proposition E cannot be known a priori, but only by empirical investigation.

His argument for (IT) then proceeds in two stages. First, he tries to show that the dependence of mental content on externalistic facts mentioned in (B) holds by conceptual necessity, i.e., that the externalistic fact upon which a given thought depends must be self-evidently inferable from either that thought itself or that thought in addition to truths known only a priori. He then goes on to argue that if this dependence is conceptually necessary, then (A)-(C) are inconsistent, thus that Privileged Access is incompatible with externalism.

### 1a. The First Stage

Before discussing the first stage of his argument, we should make explicit the notion of a priority McKinsey uses. He defines a priority as "knowledge obtained independently of empirical investigation;"[2] while this is vague, it is no more vague than the notion of non-empiricality I appealed to in Chapter II when characterizing Privileged Access, and I'll accept it at face value for now.

---

[2] McKinsey (1991), 9.

As a preliminary step, McKinsey claims that if the notion of necessity involved in (B) is merely metaphysical, then (A)-(C) are clearly consistent. His reason is that "metaphysical dependencies are often only knowable a posteriori."[3] McKinsey illustrates by drawing from Saul Kripke's (1980) argument that one's existence metaphysically entails the existence of the zygote from which one developed:

> Oscar might know a priori that he exists, and his existence might metaphysically depend upon the existence of his mother, even though Oscar cannot know a priori that his mother exists.[4]

McKinsey's idea seems to be that for all Oscar knows, let's say, he might be a clone or one of Davidson's swampmen, and he can know that he isn't such a creature only by empirical investigation. By parity of reasoning, McKinsey suggests that Oscar can't know a priori that his thought depends on E, because such a dependence is only knowable a posteriori.

---

[3] Ibid., 13.

[4] Ibid. Of course, one may not agree that the existence of a person metaphysically depends on the existence of her mother; but the example is merely intended as an illustration.

Unfortunately, McKinsey's reasoning is invalid. Just because some metaphysical dependencies can't be known a priori doesn't mean that no metaphysical dependencies can be known a priori, and McKinsey's offered no reason at all for thinking that the metaphysical dependence of thoughts on externalistic facts can't be known a priori. This will become important later on in the chapter.

Despite the invalidity of his reasoning here, though, let's grant for the moment that if mental content depends only metaphysically on external factors, then (IT) is false. This would mean that all those who think both that externalism is a metaphysical thesis and that it is incompatible with Privileged Access hold inconsistent beliefs, and one belief or the other would have to be given up. This explains McKinsey's motives in pursuing the style of argument he pursues. He retains the intuition that externalism is incompatible with Privileged Access, and opts to give up the former belief.

What's his argument against the claim that externalism is a metaphysical thesis, then? He argues that if the dependence of mental content on external factors is merely

metaphysical, then externalism is a philosophically trivial thesis, and takes for granted that externalism, whatever its content, is not philosophically trivial.

This is an amazing claim. Why should we believe it?

To see why McKinsey thinks we should, let's return to the example of Oscar and his mother. If Oscar's existence depends metaphysically on his mother's, then the existence of anything which entails his existence also depends on his mother's. So if externalism says that mental content depends metaphysically on external factors, that Oscar's existence depends on his mother's entails externalism:

> While it may be true that Oscar's thinking that water is wet entails the existence of Oscar's mother or the existence of the egg from which Oscar developed, it would nevertheless not be for this reason that Oscar's mental state is wide![5]

Well, not exactly. The entailment obtains only if the existence of Oscar's mother is an externalistic fact; why does McKinsey think it is? He thinks so because he defines externalism as the thesis that

---

[5] Ibid., 14.

> [s]ome neutral cognitive states that have propositional content depend upon or presuppose the existence of objects external to the persons in those states.[6]

It's worth noting here that this definition of externalism is quite different from the one offered in Chapter III, and this alone is reason to suspect that something is wrong with McKinsey's argument. This is an important point, but one I will postpone developing until the end of this section.

Given McKinsey's definition of externalism, then, it's easy to see why McKinsey thinks the existence of Oscar's mother is an externalistic fact: Oscar's mother is external to Oscar, and Oscar's thinking that water is wet--a neutral cognitive state-token of Oscar's--depends on Oscar's existence, hence on the existence of his mother. If this is correct, then if externalism, defined as above, is a metaphysical thesis, it follows trivially from metaphysical theses which are independent of particular semantic theories.[7] For example, externalism so understood is

---

[6] Ibid.

[7] It may not be correct, however, because 'neutral cognitive state' in McKinsey's definition of externalism may also refer to propositional-attitude types instead of the token states that realize them. As we'll see shortly, McKinsey's argument depends on 'neutral cognitive state' referring to state tokens. See section 1c.

compatible with both indexical-description and conceptual-role theories of content. As externalism is clearly not consistent with such theories, it seems the dependence of mental content on external factors can't hold merely by metaphysical necessity. The dependence must be stronger:

> Clearly, to say that [Oscar's thought that water is wet] is wide is not to say something that is true by virtue of Oscar's nature or the nature of the particular event that is Oscar's thought that water is wet. Rather it is to say something about the concept, or property, that is expressed by the English predicate 'x is thinking that water is wet'; it is to say something about what it means to say that a given person is thinking that water is wet.[8]

That is, externalistic entailments such as (B) are known analytically by those who entertain them: any rational speaker with the concepts expressed in (B) knows on understanding it that it is true. In this way, McKinsey concludes that the dependence of mental contents on external factors holds by conceptual necessity.

### 1b. The Second Stage

This stage of the argument is straightforward. Recall that for a dependence of x on y to hold by conceptual

---

[8] Ibid.

necessity is for y to be self-evidently inferable from either x alone or x plus only truths known a priori. What does this mean? McKinsey defines it as follows:

> Let us say that a proposition p conceptually implies a proposition q if and only if there is a correct deduction of q from p, a deduction whose only premisses other than p are necessary or conceptual truths that are knowable a priori, and each of whose steps follows from previous lines by a self-evident inference rule of some adequate system of natural deduction.[9]

His notion of conceptual implication leaves it open that some premises in a deduction of q from p are non-conceptual, that is, presumably non-analytic. But because he has it that externalistic entailments are known analytically--on the basis of what they mean--the only non-analytic premise appealed to in the deduction of externalistic consequences will be the attributions of self-knowledge to thinkers.

Now, if the dependence of mental content on external factors holds analytically, then if one knows a priori what one is thinking, and one's thought depends on an externalistic fact analytically, then one's thought and the analytically known externalistic consequence together conceptually imply that consequence. So, to return to (A)-

---

[9] Ibid.

(C), if (A) and (B) are true and the dependence in (B) is known analytically, then Oscar can infer E by Modus Ponens from the proposition that he is thinking that water is wet and the proposition that that Oscar is thinking that water is wet entails E. But this clearly conflicts with (C), which says that E can only be known by means of empirical investigation. Therefore, (A)-(C) are inconsistent, and so Privileged Access and externalism are incompatible.

### 1c. Critique of McKinsey

Clearly, the force of McKinsey's argument for (IT) depends on the force of his argument that externalistic entailments are known analytically. In this section, I show that his argument is based on a conflation of beliefs with the states which realize them; when we see that externalism is a thesis about beliefs alone, he cannot argue for thesis by depending on a thesis about the conceptual dependence of realizing states on externalistic facts.

It should also be clear that the force of McKinsey's argument for the claim that externalistic entailments hold analytically depends on the force of his argument that if the dependence in (B) is merely metaphysically necessary,

then externalism is a philosophically trivial thesis. How good is this argument?

At this point, it's worth restating McKinsey's definition of externalism:

> Some neutral cognitive states that have propositional content depend upon or presuppose the existence of objects external to the persons in those states.[6]

Let's understand the dependence in this definition as metaphysically necessary, so that it is a philosophically trivial thesis. Let's call this thesis McKinsey Externalism.

Is McKinsey Externalism trivial? Well, it depends on what one means by 'neutral cognitive state' in the definition. If a neutral cognitive state is a propositional attitude property, then McKinsey Externalism is clearly not trivial, although not precisely equivalent to the definition of externalism given in Chapter III.[10] Oscar's clone or Davidson's Swampman can believe that water is wet, even

---

[10] Externalism as defined in Chapter III makes reference to factors external to a thinker; these factors may or may not involve the existence of objects external to thinkers, depending on the strength of externalism under discussion.

though neither has a mother.[11] Indeed, there are logically possible worlds in which the sole realizer of propositional attitude properties is a single creature existing eternally. On the other hand, if a neutral cognitive state is a propositional attitude token, such as a neural state of Oscar's, clearly some neutral cognitive states presuppose the existence of Oscar's mother. If we understand McKinsey's definition in this way, McKinsey Externalism really is trivial.[12]

How should we understand McKinsey Externalism, then? We noted above that McKinsey Externalism, understood as referring to the token realizers of propositional-attitude properties, is compatible with even individualistic theories of content, such as indexical-description and conceptual-role theories. (Call this Token McKinsey Externalism.) It's easy to see why. Suppose, for the sake of argument,

---

[11] See Davidson (1987) for the Swampman example. If Oscar's clone or Swampman believe that Oscar is thinking that water is wet, their believing it presupposes Oscar's mother's existence, but only supposing that the proposition presupposes Oscar's existence and that Oscar's existence presupposes his mother's. Only the second supposition is independent of any particular theory of content.

[12] I am grateful to Stephen Schiffer for helpful discussion of this point.

that it's merely metaphysically necessary that molecule-for-molecule identical thinkers have all the same beliefs, desires, etc. Perhaps this is because a mental state's conceptual role is metaphysically sufficient for determining its content, perhaps for some other reason; but in effect, we're supposing that the following conditional is false:

> If x is thinking that water is wet, it's metaphysically possible that a creature molecule-for-molecule identical to x is not thinking that water is wet.

The problem is that even if this is so, Token McKinsey Externalism is still true! That's because even if Oscar and Oscar's twin share all their beliefs by metaphysical necessity, that they have intentional state (tokens) at all depends on the existence of their mothers. So even if the propositional contents of one's thoughts can't vary from environment to environment, Token McKinsey Externalism is still true.

This is odd. Were Putnam and Burge so daft as to argue for a thesis that clearly doesn't entail the sorts of claims they think it does? Perhaps McKinsey would say No, but conclude that they must have instead been arguing for the thesis that thoughts conceptually depend on externalistic

factors. Unfortunately, this doesn't follow. All that follows is that they weren't arguing for Token McKinsey Externalism. They could well have been arguing for the thesis that it's metaphysically possible for twins to differ in their beliefs, and this thesis clearly isn't equivalent to the thesis that thoughts conceptually depend on externalistic facts. In fact, the thesis that it's metaphysically possible for twins to have distinct beliefs is equivalent to the definition of externalism given in Chapter III, i.e., the thesis that

> mental content is typically, if not always, determined by factors external to thinkers.

So the token version of McKinsey Externalism is a red herring. Token McKinsey Externalism is trivial, but it's simply not externalism. Externalism is a thesis about the dependence of propositional-attitude properties on external factors, not a thesis about the dependence of realizers of propositional-attitude properties on external factors.

No doubt, many who discuss externalism do so sloppily, and in ways that make it indistinguishable from such trivial theses as Token McKinsey Externalism. But if we keep in focus the actual targets of those who defend Externalism--

individualistic theories of content--then it should be clear that what's at issue is the metaphysical dependence of propositional-attitude properties on external factors, not the dependence of their realizing token states.

### 1d. Conclusion

Now, it doesn't follow from my critique that, on McKinsey's definition of conceptual implication, Oscar's thinking that water is wet doesn't conceptually imply its externalistic consequence if externalism is true. Of course, it should be clear that externalism in no way implies that Oscar's thinking that water is wet analytically implies any externalistic consequence. Metaphysical dependencies needn't be known at all, much less known analytically. But McKinsey's notion of conceptual implication doesn't rule out the possibility that metaphysical dependencies can be known a priori, which means that the argumentative strategy McKinsey uses may yet yield a plausibly sound argument. In the fourth section of this paper, I discuss the question of whether externalism in fact implies that people can know externalistic consequences on the basis of McKinsey-style conceptual implication.

None of this changes the present situation, however: that McKinsey has failed to establish (IT). His argument for it falters at the first stage by mischaracterizing externalism, because his argument that externalistic entailments hold by conceptual necessity is based on a definition which conflates propositional-attitude properties with their realizers; because externalism is a thesis about properties alone, the second stage of the argument shows that one can know empirical facts a priori only if an exceedingly implausible version of externalism is true.

## 2. Brueckner's Argument

In contrast to McKinsey's highly involved argument for (IT), Brueckner's (1990) argument is simple. Brueckner argues that if externalism is true, then a version of a well known skeptical argument shows that we lack privileged access to our own mental states. His argument applies this skeptical strategy to an externalistic framework, so it's worth taking a brief look at this strategy before discussing his argument; I do this in the next part of this section. In the next two parts, respectively, I develop Brueckner's argument and critique it. As we'll see soon, like

McKinsey's argument, Brueckner's argument also falters at an early stage.

## 2a. Skepticism and Closure

The idea behind the standard skeptical argument against knowledge of the external world is to show that we are in no position to eliminate possibilities which we know to be incompatible with what we claim to know about the external world. So, for example, I know that if I have two hands, then I am not a handless brain in a vat. But because I'm not in a position to rule out the possibility that I am a handless brain in a vat, it seems to follow that I don't know that I have two hands.

This reasoning implicitly relies on the following prima facie plausible closure principle:

If I know both p and that if p, then q, I can know q.

Let's call this principle Closure.[13]

---

[13] The way of stating Closure that's standard in the literature leaves out the 'can', but such a principle, understood literally, is clearly false. I might know p and that if p then q, but not have gotten round to drawing the inference. The sense of 'can' here is straightforward: I can know q in that I have the capacity to draw the inference. We should understand standard Closure to involve the implicit 'can' for charity's sake.

By relying on this principle, it can be shown that if I'm in no position to rule out the possibility that I'm a handless brain in a vat, then I don't know that I have two hands. The argument for the claim is simple and compelling:

> Let's suppose that I <u>am</u> a handless brain in a vat. In such a situation, everything would seem to me exactly as it seems now. But if that's so, then my evidence doesn't rule out the possibility that I am a handless brain in a vat: if my evidence <u>did</u> rule it out, then there would be some feature of my experience which could not be equally well explained by the hypothesis that I am a handless brain in a vat. What's more, if I can't rule out the possibility that not-p, then I don't know p. So I don't know that I'm <u>not</u> a handless brain in a vat. But I know that if I have two hands, then I'm not a handless brain in a vat. Therefore, I don't know that I have two hands.

This strategy can be used to refute just about any interesting knowledge claim about the external world; just substitute it for the claim that I have two hands.

One way of responding to this argument is to <u>deny</u> Closure: Fred Dretske (1970) and Robert Nozick (1981) take this line. I will discuss this line in connection with the paradox in the next chapter, when I discuss previous solutions to the paradox; for now, though, let's accept Closure.

Besides Closure, the above argument's force also depends on the plausibility of the principle that if one is in no position to rule out not-p, then one doesn't know p. The idea behind this is that one's evidence for the proposition one claims to know must rule out the counter-possibility raised by the skeptic; that is, if I know that I am not a handless brain in a vat, then my evidence must favor the hypothesis that I have hands over the hypothesis that I am a handless brain in a vat. Thus, when it comes to empirical knowledge at least, to say that I am in a position to rule out not-p just is to say that my evidence favors p over not-p.[14]

Of course, there are indefinitely many propositions incompatible with any proposition one knows; so if both Closure and the above principle hold, then if I know that I

---

[14] I am supposing for the sake of discussion that evidence E favors p over q iff p but not q is part of the best explanation of E. The proper relation of evidence to skeptical hypotheses is unclear, however, and I won't pretend to settle the issue here.

Also, one can be in a position to rule out not-p even though one possesses no evidence at all with respect to p--privileged access being an apparent case in point. So one can be in a position to rule out not-p by one's belief p being self-warranted.

have two hands, I am in a position to rule out every proposition which implies that I lack at least one. This may seem too strong, however: some believe that one needn't be in a position to rule out every proposition incompatible with what one knows, that in some relaxed circumstances one can know p even when one's evidence doesn't favor p over some incompatible proposition q. This view, which is known as contextualism, is defended by Stewart Cohen (1988), Keith DeRose (1995), and David Lewis (forthcoming). I will also discuss contextualism in the next chapter, but for now, let's accept the principle it challenges.

It's easy to see this skeptical strategy leads to a paradox consisting of the following three proposition schemas, where p is a proposition about the external world and q a skeptical proposition which I know to be incompatible with p:

(A1) I know p.

(A2) If I know p, I know not-q.

(A3) I don't know not-q.

Call this the External-World Paradox. The External-World Paradox has received much attention of late (see, for

example, Barry Stroud 1983, Michael Williams 1991, and DeRose 1995), but I will have little to say about it here. My concern is the application of this paradox to the case of externalism and Privileged Access. Brueckner thinks that if externalism is true, then the External-World Paradox can be extended to include propositions ascribing mental states to oneself. Let's see how his argument works as an attempt to demonstrate (IT).

### 2b. Applying Skepticism to Privileged Access

To make his case for skepticism about privileged access, Brueckner imagines a scenario in which some person we'll call Sally is unknowingly transported from her native Earth to Twin Earth. Sally notices no difference at all, of course. After a time, Sally's utterances of such sentences as 'Water is wet' and occurrent 'water'-thoughts will no longer express propositions about water; instead, they will express propositions about twater. Nevertheless, Sally at no time becomes aware of the shift in the contents of her thoughts and utterances.[15]

---

[15] Boghossian (1989, 13) points out that another way of describing switching scenarios is possible. By virtue of

Given this scenario, Brueckner applies the reasoning behind the External-World Paradox to generate a parallel skeptical paradox about one's own mental states:

> I claim to know that I am thinking that some water is dripping. If I know that I am thinking that some water is dripping, then I know that I am not thinking, instead, that some twater is dripping. But I do not know that I am not thinking that some twater is dripping, since, according to externalism, if I were on twin earth thinking that some twater is dripping, things would seem exactly as they now seem (and have seemed). So I do not know that I am thinking that some water is dripping.[16]

---

slow switching, it may be that one can acquire the concept twater without thereby losing the capacity to express thoughts about water. (He argues that a particularly strong version of this claim is actual in Boghossian 1992.) In this scenario, he notes (note 11) that "there is no simple answer" to the question of which proposition one expresses in uttering sentences such as 'Water is wet' and in having the correlative thoughts. Without further consideration, it would be premature even to rule out the possibility that there is no fact of the matter as to which proposition is expressed. For example, it may be that in switching scenarios, 'water' partially denotes both water and twater, as in Hartry Field (1974). (I am grateful to Stephen Schiffer for helpful discussion of this point.)

It's worth noting that if there is no fact of the matter as to which proposition one expresses in having thoughts in switching scenarios, then there is a simple argument from externalism to the denial of privileged access, at least in the case of switching subjects: because there is no fact of the matter about what one is thinking, one cannot know, empirically or non-empirically, what one is thinking--for knowing is factive.

[16] Brueckner (1990), 448.

Like the other skeptical argument, the force of this one seems to depend on Closure and the principle that if one is in no position to rule out not-p, one does not know p. It can be represented more perspicuously as follows:

(B1) I know that I am thinking that some water is wet.
(B2) If I know that I am thinking that some water is dripping, I know that I am not thinking that some twater is dripping.
(B3) I don't know that I am not thinking that some twater is dripping.

(B1)-(B3) are instances of (A1)-(A3), substituting 'that I am thinking that some water is dripping' for 'p' and 'that I am thinking that some twater is dripping' for 'q'. But there are important differences. First, (A1) is taken to be true without qualification, while (B1) is supposed to be a consequence of privileged access. Second, (A3) is supposedly a consequence of skeptical reasoning alone, while (B3) presupposes not only such reasoning, but also both externalism and my lacking <u>empirical</u> knowledge that water and twater differ.[17] (If I lack empirical knowledge of the

---

[17] Brueckner (ibid.) notes that if I have empirical knowledge of the difference between water and twater, I can infer that my water-thoughts and twater-thoughts are different from the fact that my empirical investigations show my thoughts to be about distinct substances. He

difference between water and twater, there is no introspectible difference between thoughts about them.)

So (B1)-(B3) is an application of (A1)-(A3) only given externalism, privileged access, and my lacking certain empirical knowledge. But this is just as well, because Brueckner is trying to motivate the paradox of externalism and privileged access, and so is implicitly offering an argument for (IT). Given these considerations, then, we can embed (B1)-(B3) into the following larger argument for (IT),

---

correctly points out that this is the wrong kind of knowledge, because it "discredits the intuitively plausible idea that I can know the contents of my thoughts directly, without relying on ordinary empirical knowledge...." Unfortunately, he also seems to assume that having such empirical knowledge in itself rules out the possibility of Privileged Access altogether.

Such an assumption is hard to support. If I have empirical knowledge of the difference between water and twater, then my concepts of the two have distinct roles in my inferences. So if I have privileged access to the conceptual roles of my thoughts, my empirical knowledge of the difference between water and twater no more undermines my privileged access to water thoughts any more than does my empirical knowledge of the difference between water and Frank Sinatra.

Brueckner seems to be relying on the idea that the concepts water and twater are introspectively indistinguishable for switching subjects. This entails that the two play identical roles in inference. Therefore, if they play distinct conceptual roles, like water and Frank Sinatra, they ought to be introspectively distinguishable.

assuming that I am thinking that some water is dripping and lack knowledge of the difference between water and twater:

(C1) If we have privileged access, then I know introspectively that I am thinking that some water is dripping.

(C2) If externalism is true, then if I am thinking that some water is dripping, I am not thinking that some twater is dripping.

(C3) I know introspectively that if I am thinking that some water is dripping, then I am not thinking that some twater is dripping.

(C4) If I know introspectively both p and that if p, then q, I can know q introspectively.

(C5) But if externalism is true, then I don't know introspectively that I am not thinking that some twater is dripping.

(C6) Therefore, if externalism is true, I don't know introspectively that I'm thinking that some water is dripping.

(IT) Therefore, privileged access and externalism are incompatible.

(B1) appears as the consequent of (C1), because it depends on privileged access; (B2) is represented as (C3); and (B3) appears as the consequent of (C5), because it depends on both externalism and the assumption underlying the argument. Therefore, Brueckner's application of the External-World

Paradox, (B1)-(B3), does play a role in an argument for (IT), but only in the context of (C1)-(C6).

### 2c. Critique of Brueckner

The argument from (C1)-(C6) to (IT) is valid. To assess its soundness, then, we must examine its premises--(C1)-(C5). We can safely ignore (C1) and (C2), for they are clear consequences of externalism and Privileged Access[18]; let's also ignore (C4), because for the sake of argument we're not challenging Closure, and (C4) is a highly plausible version of Closure. (It's plausible that if Closure holds, then (C4) also holds.) This allows us to restrict our examination to (C3) and (C5).

Neither (C3) nor (C5) follow from either externalism or Privileged Access, although each might be true independently of either thesis. The case for (B5) is straightforward: on the assumptions that the consequent of (C2) is true and that I lack knowledge of chemistry, (C5) seems to follow. The reason is that on those assumptions, I either don't have a

---

[18] Of course, (C2) is a clear consequence of externalism only if we assume that Brueckner's description of switching scenarios exhausts the possibilities; but we are assuming for the purposes of discussion that it does.

concept of twater at all, or that my concept of water isn't distinct from my concept of twater.

What is a lack of distinctness of concepts? Sally is a good example of someone who has two concepts which fail to be distinct. Having been secretly transported to Twin Earth, she lives there long enough for her 'water'-thoughts and 'water'-utterances to acquire the content twater. Of course, she still has the concept water--if she were to be returned to Earth and subsequently utter the sentence, "Water is a liquid," she would correctly be taken to have said that water is a liquid. So Sally has both the concept water and the concept twater. But these concepts aren't distinct: she can't entertain the thought that something falls under one concept but not the other.

Given that I know that I'm thinking that some water is dripping, that I lack knowledge of chemistry, and have never spent time on Twin Earth, it should follow that I don't know that I'm not thinking that some twater is dripping. But if I have spent enough time on Twin Earth to acquire the concept twater, I still wouldn't know that I'm not thinking that some twater is dripping, because I wouldn't be able to

entertain the conjunctive proposition that I'm thinking that some water is dripping, but not that some twater is dripping. If I could even think simultaneously both that I'm thinking that some water is dripping and that I'm not thinking that some twater is dripping, I would be able to distinguish between water and twater, and so would have distinct concepts. So if my concepts water and twater are not distinct and I know that I'm thinking that some water is dripping, I don't know that I'm not thinking that some twater is dripping. So (C5) seems true.

Unfortunately, the case we've made for (C5) clearly falsifies (C3)! Here's why. First, if I lack the concept twater altogether, then I don't know that if I'm thinking that some water is dripping then I'm not thinking that some twater is dripping--I don't know it just because I don't even believe it. (I can't have such beliefs because I lack the concept twater.) Second, if I have the concept twater but not distinctly, then I also don't know that if I'm thinking that some water is dripping then I'm not thinking that some twater is dripping, because I don't even believe that proposition either--if I could believe it, my concepts

would be distinct. Finally, if I have the concept twater distinctly, then I do know that if I'm thinking that some water is dripping then I'm not thinking that some twater is dripping; unfortunately, in such a scenario (C5) is clearly false! (If it's not clear why, just substitute a distinct concept for twater, say, alcohol: the substituted version of (C3) will be true and that of (C5) will be false.)[19]

Brueckner's argument, then, clearly can't be plausibly sound, because whatever case can be made for one of its premises undermines the plausibility of the other.

### 2d. Conclusion

Once again, just because Brueckner's argument for (IT) doesn't work doesn't mean that no argument skeptical of Privileged Access can be culled from externalism. Later on, we'll see if improvements are possible. But for now, we should see that the argument Brueckner actually offers for (IT) fails, because he can't plausibly maintain both (C3) and (C5), but needs both for his argument.

---

[19] I am indebted to Schiffer here for help in sorting out this point. Schiffer has made virtually the same point in his lectures, although he didn't use the notion of concept distinctness.

## 3. Boghossian's Argument

Boghossian's argument for the Incompatibility Thesis is much like Brueckner's in that it, too, is a direct argument from externalism to skepticism about Privileged Access. But it differs in an important way: whereas Brueckner's argument depends on Closure, Boghossian's does not. Nevertheless, we'll see that despite this difference, the argument fails for reasons quite similar to Brueckner's.

### 3a. ("Inner") Perception and Closure

Boghossian is tempted to characterize Privileged Access in terms of an "inner awareness" of one's own mental states:

> The suggestion that I know about my thoughts by being introspectively aware of them seems, from the phenomenological standpoint anyway, overwhelmingly plausible. It is not simply that I have reliable beliefs about my thoughts. I catch some of my thoughts in the act of being thought. I think: If she says that one more time, I'm leaving. And I am aware, immediately on thinking it, that that is what I thought.[20]

But he thinks a problem arises when we consider Privileged Access in light of externalism: if externalism is true, mental content is determined by relational properties:

---

[20] Boghossian (1989), 11.

> [H]ow could anyone be in a position to know his thoughts merely by observing them, if facts about their content are determined by their relational properties?[21]

Precisely what the problem is, though, is not clear; so Boghossian uses an analogy with perception to clarify it.

Suppose I claim to know, by looking, that there is a dime in my hand. A skeptic points out that I know full well that if I have a dime in my hand, then I do not have a cunning counterfeit in my hand. But (let's suppose that) I clearly don't know that I don't have a cunning counterfeit in my hand--I can't distinguish dimes from cunning counterfeits. Part of what makes a dime a dime is indeed something relational--its having been created by the U.S. Mint. Because I am in no position to claim that the coin in my hand was definitely created by the U.S. Mint, the skeptic concludes, I don't really know I have a dime in my hand.

This skeptical argument exploits the fact that in ordinary cases of perceptual knowledge we rarely, if ever, know that the conditions which make our claims true in fact obtain. Now, it's central to externalism that thinkers can count as having the thoughts they actually have without

---

[21] Ibid.

knowing that the conditions which make them possible hold. So why doesn't this argument apply to our own thoughts?

We've already noted one serious problem with applying this argument to the case of our own thoughts: it's simply not true that one in general knows such conditionals as If one is thinking that water is a liquid, then one is not thinking that twater is a liquid, and such conditionals are essential to applying Closure to the case of Privileged Access. But Boghossian notes another problem:

> Someone may know, by looking, that he has a dime in his hand. But it is controversial, to put it mildly, whether he needs to know all the conditions that make such knowledge possible. He need not have checked, for example, that there is no counterfeit money in the vicinity, nor does he need to be able to tell the difference between a genuine dime and every imaginable counterfeit that could have been substituted.[22]

So even if I know that if I have a dime in my hand, then I do not have a cunning counterfeit in my hand, my failure to know that I do not have a cunning counterfeit doesn't seem to undercut my claim to know that I have a dime in my hand. I certainly ought to be able to tell the difference between a dime and other coins, legitimate and counterfeit. Such

---

[22] Ibid., 12.

alternatives are those one would expect me to know don't obtain if I'm to legitimately claim to know that I have a dime in my hand: they are, to use Fred Dretske's (1970) term, relevant alternatives. That I have a cunning counterfeit in my hand is not one of these.

If Boghossian is right, then the Closure principle which underlies skeptical arguments doesn't apply to ordinary perceptual knowledge; and because he understands Privileged Access to be akin to such knowledge--it is "inner" as opposed to "outer" awareness, let's recall--he thinks it's doubtful that skeptical arguments against Privileged Access can legitimately appeal to Closure. Thus, just because one doesn't know that one is not thinking that twater is a liquid doesn't imply that one doesn't know that one is thinking that water is a liquid, even if one knows the relevant conditional. The possibility that one is thinking that twater is a liquid is not a normally relevant alternative.

### 3b. Relevant Alternatives and Switching Cases

So what blocks Closure, according to Boghossian, is that "the ordinary concept of knowledge appears to call for

no more than the exclusion of 'relevant' alternative hypotheses,"[23] and that the alternatives offered in skeptical arguments are often only logically possible. And mere logical possibility is not sufficient for relevance.

Nevertheless, Boghossian thinks that the "too swift" argument offered in 3a "suggests a slower and more convincing argument for the same conclusion,"[24] because

> it seems easy to describe scenarios in which the twin hypotheses are relevant alternatives, but in which they are, nevertheless, not discriminable non-inferentially from their actual counterparts.[25]

The scenarios he has in mind, of course, are the very switching cases Brueckner describes:

> Imagine that twin-earth actually exists and that, without being aware of it, S undergoes a series of switches between earth and twin-earth.... [W]e may imagine that after a series of such switches, S ends up with both earthian and twin-earthian concepts: thoughts involving both [water] and [twater] are available to him.[26]

Thus, the possibility that Sally is thinking that twater is a liquid is a relevant alternative for her, because she has

---

[23] Ibid.

[24] Ibid., 13.

[25] Ibid.

[26] Ibid.

both the concept water and the concept twater. This leads to an argument very like Brueckner's:

(D1) If externalism is true, then if I'm thinking that water is a liquid, I'm not thinking that twater is a liquid.

(D2) If I have privileged access, then I can discriminate non-inferentially between all the relevant alternatives to my actual thoughts.

(D3) If I am a switching subject, then the possibility that I'm thinking that twater is a liquid is a relevant alternative.

(D4) If the possibility that I'm thinking that twater is a liquid is a relevant alternative, then if I know non-inferentially that I'm thinking that water is a liquid, I know non-inferentially that I'm not thinking that twater is a liquid.

(D5) I don't know non-inferentially that I'm not thinking that twater is a liquid.

(D6) Therefore, if externalism is true, I don't know non-inferentially that I'm thinking that water is a liquid.

(IT) Therefore, privileged access and externalism are incompatible.

### 3c. Warfield's Critique

Ted Warfield (1992) noted that the argument (D1)-(D6) is invalid: (D6) doesn't follow from (D1)-(D5). What does follow is the much weaker

(D6') If I am a switching subject, then I don't know non-inferentially that I'm thinking that water is a liquid.

But (IT) doesn't follow from (D1)-(D6'); so Boghossian's argument has a gaping hole in it.

Warfield is absolutely correct. The only problem with his criticism is that it makes Privileged Access contingent. It's odd to suppose someone can know what she's thinking while her twin doesn't, because whether or not one knows what one is thinking doesn't seem to depend on environmental or social factors.

Still, what one feels about the contingency of Privileged Access in no way affects Warfield's point. Boghossian's argument is patently invalid. Can we fix it?

No doubt, Boghossian tacitly relies on the intuition that introspective knowledge of one's thoughts does not depend on whether one is a switching subject. Perceptual knowledge does depend on the character of the perceiver's environment, but why should privileged access? To suppose that one's knowledge of one's mental states does depend on

the character of one's environment seems an admission that such knowledge is not (merely) introspective.[27]

This suggests that we can fix Boghossian's argument by replacing (D6) with (D6') and adding the following premises:

(D7) If I know introspectively that I am thinking that water is a liquid, then my knowledge that I am thinking that water is a liquid doesn't depend on the character of my environment.

(D8) If (D6'), then my knowledge that I am thinking that water is a liquid does depend on the character of my environment.

(D9) Therefore, I don't know introspectively that I am thinking that water is a liquid.

The argument from (D1)-(D6')-(D9) to (IT) is valid, thus avoiding Warfield's charge of invalidity.[28]

---

[27] Warfield asserts without argument (1992, 235) that an externalist can happily accept that a thinker's introspective knowledge of her own mental states is a contingent matter. It should be clear from this discussion that the externalist's happiness must be earned.

[28] Peter Ludlow (1995) tries a different way of fixing Boghossian's argument. He suggests that switching cases are in fact prevalent, and that the possibility of twin thoughts is a relevant alternative just in case switching cases are prevalent, whether or not they are actual in one's own case.

I find these suggestions unhelpful. If Ludlow is prepared to accept that Privileged Access does in fact depend on the character of one's environment, then Warfield would probably grant that what makes a twin case a relevant alternative is the prevalence of switching cases. Granting that, it is then an empirical question whether any particular person does know what she is thinking--which is

### 3d. Critique of Boghossian

Having met Warfield's objection, should we suppose that Boghossian's argument is sound? In considering its soundness, we should examine the notion of relevant alternative to which Boghossian is appealing. What is a relevant alternative?

The necessary and sufficient conditions for an alternative's being relevant aren't clear.[29] But a clearly necessary condition for an alternative's being relevant is that a putative knower at least know that it's incompatible with what she claims to know. That is, if q is a relevant alternative to p, then for any knower x,

x knows that if p, then not-q.

The skeptic doesn't require that a knower be able to rule out alternatives she doesn't realize are incompatible with what she claims to know, but only those she knows are incompatible with what she claims to know. The denial of

---

precisely what Warfield claims.

[29] Indeed, they shouldn't be clear. If they were, then the External-World Paradox would have an obvious solution, which it doesn't.

Closure restricts this still further: among those alternatives a knower knows are incompatible with those she claims to know, she need rule out only those which are relevant (whatever relevancy amounts to). So both the skeptic and her opponent agree that a knower must know that an alternative is incompatible with what she claims to know.

If this is right, though, then Boghossian's argument fails for precisely the same reason that Brueckner's does. The key premise,

(D3) If I am a switching subject, then the possibility that I am thinking that twater is a liquid is a relevant alternative,

is false. If I am a switching subject, then I don't have distinct concepts of water and twater, which means that I don't know that the following conditional is true:

If I am thinking that water is a liquid, then I'm not thinking that twater is a liquid.

But if I don't know this conditional, then the possibility that I am thinking that twater is a liquid is not a relevant alternative, because that possibility's being a relevant alternative depends on my knowing this conditional.

Boghossian's argument, then, is no more plausible than Brueckner's. Both require that thinkers know conditionals which they clearly wouldn't know if externalism were true.

## 4. Improving the Arguments

We've seen that neither McKinsey's, Brueckner's nor Boghossian's arguments are plausibly sound. Nevertheless, the intuition that there is some conflict between externalism and Privileged Access really ought to lead to a prima facie sound argument, if only because so many hold it. It may be that the reason that none of the arguments for (IT) thus far offered are successful is that they are fledgling attempts to articulate a sound intuition; further examination may yield stronger arguments. I make such an examination in this section.

In the first part of this section, I examine McKinsey's argument anew, and show that a much stronger argument with at least prima facie plausibility--due to Schiffer (lectures)--is available to him. In the second part, I revive Brueckner's argument, and argue that unlike McKinsey's, it can't be made plausibly sound; that's because it depends on the truth of conditionals which simply can't

be true if externalism is true. This explanation also shows why Boghossian's argument also can't be made to work.

### 4a. Improving McKinsey

Let's recall McKinsey's notion of conceptual implication:

> Let us say that a proposition p conceptually implies a proposition q if and only if there is a correct deduction of q from p, a deduction whose only premisses other than p are necessary or conceptual truths that are knowable a priori, and each of whose steps follows from previous lines by a self-evident inference rule of some adequate system of natural deduction.[7]

We noted that McKinsey argued that the externalistic entailments such as those in (B) held by virtue of what such predicates as 'x is thinking that water is wet' mean--that is, that anyone who understands these predicates knows (B)-style entailments. But we also noted that McKinsey's notion of conceptual implication doesn't require that conceptual implications be drawn on the basis of analytically known meaning relations: we can also draw conceptual implications by appeal to necessary truths which are knowable a priori.

Of course, we can't rely too much on the notion of a priority; but it seems clear that if externalism is true,

then there are at least some necessary truths which are knowable a priori without being known analytically. For example, externalism itself is a thesis that, if true, is true by metaphysical necessity; but the arguments for externalism neither appeal to particular empirical facts nor depend for their force on the weight of empirical evidence. So if one understands such arguments and sees that they have this non-empirical aspect, then one can know them to be true on the basis of these arguments alone: that is, one can know them to be true a priori, in the sense of 'a priori' McKinsey himself uses.

Now, this is a rather controversial point, especially because it makes use of the unhappy notion of a prioricity. I'll discuss this notion in more detail in Chapter VI. For the sake of discussion, though, let's accept it.

In any event, it should also be clear that anyone who knows externalism a priori should also know a priori the externalistic entailments which are logical implications of externalism. For example, externalism implies that if one is thinking that water is wet, then some externalistic fact E obtains--the exact nature of this fact depending on the

particular variety of externalism one favors. So if one knows, by knowing logic, that externalism logically implies this conditional, and one knows externalism a priori, then one ought to be able to know this conditional a priori.

The kind of reasoning described by McKinsey-style conceptual implication, then, is a kind of Closure we may call a priori Closure:

> If x knows a priori both p and that if p then q, then x can know q a priori.

A priori Closure is very plausible: logical and mathematical reasoning particularly seem to depend on it. We noted very early in this chapter that the notion of a priority McKinsey appeals to is very similar to the notion of non-empiricality I appealed to in Chapter II. So we might also suppose that if a priori Closure holds, so too does its non-empirical analogue:

> If x knows non-empirically both p and that if p, then q, x can know q non-empirically.

We'll examine in Chapter V whether such Closure principles in fact hold, despite their intuitiveness. For now, though, let's accept them.

Given these principles, the following argument that externalistic entailments can be known non-empirically emerges:

(E1) That externalism is true logically implies conditionals of the form If x is thinking p, then E, where E is an appropriate externalistic content-determiner.

(E2) Some people know non-empirically, that externalism is true.

(E3) These same people know logic well enough to know a priori, hence non-empirically, that (E1) is true.

(E4) If x knows non-empirically both p and that if p, then q, then x can know q non-empirically.

(E5) Therefore, these people can know non-empirically conditionals of the form If x is thinking p, then E.

This argument is valid, and it seems to be sound. Two premises are clearly controversial: (E2), because externalism is a controversial philosophical thesis, and (E4), because Closure is challengeable and the plausibility of non-empirical Closure may depend on it. But we may agree that externalism, if true, is necessarily true, and it seems clear that the arguments for it don't depend on particular empirical facts; so if being a priori amounts to being knowable in a way that doesn't depend on empirical facts,

then we can provisionally agree that externalism is knowable a priori.[30] And because we're accepting Closure for the time being, we may also accept its non-empirical analogue.

If we accept this argument, then McKinsey simply doesn't need to argue that externalism is a trivial thesis if the dependence in (B) is only metaphysically necessary, or that externalistic entailments are knowable solely on the basis of the meanings of such predicates as 'x is thinking that water is wet'. All he needs as support is the above argument to establish his conclusion.

This, then, is the Improved McKinseyan Argument (IMA), where the philosopher of language Nathan Salmon is a suitably sophisticated thinker:[31]

(F1) If we have privileged access, then Nathan Salmon knows non-empirically that he is thinking that water is wet.

(F2) If externalism is correct, then Nathan Salmon knows non-empirically, that his thinking that water is wet entails externalistic fact E (where E is an appropriate externalistic content-determiner).

---

[30] But see my discussion in Section 5.

[31] For no reason other than his being a well known externalist, Nathan Salmon will serve throughout this essay as an example of a suitably sophisticated thinker. Fame has its price.

(F3) If x knows non-empirically both p and that if p then q, x can know q non-empirically.

(F4) Therefore, if we have privileged access and externalism is correct, then Nathan Salmon can know E non-empirically.

(F5) But no one can know E non-empirically.

(IT) Therefore, privileged access and externalism are incompatible.[32]

IMA, then, is our first serious contender for a plausibly sound argument for (IT).

A clarification about non-empiricality is necessary to avoid trouble later on.[33] To say that someone knows something non-empirically, as I am using the term, is to say something about the character of one's justification for the thing, not to say something about the nature of the proposition known or the way in which the proposition came to be known. (It's not the happiest term: perhaps 'non-evidential' would be better, but it has problems of its own

---

[32] This argument, slightly altered, is due to Stephen Schiffer (lectures).

[33] This clarification arose out of a conversation with Richard Mendelsohn, who pointed out that on a very reasonable sense of 'empirical', it's ludicrous to think that anyone can know externalistic consequences non-empirically. I argue later that we can have non-empirical knowledge (of a very specific sort) of externalistic consequences, so the clarification will be useful later.

I'd prefer to avoid.) But on my use of 'non-empirical', it's left open the kind of proposition one knows and how one came to know it. It may be that the kind of proposition one knows or how one came to know it somehow determines the character of one's justification for believing it, but that's a different matter.

Thus, (F5) has to do with the non-empiricality of Nathan Salmon's justification for believing E, not with the character of E itself or how Nathan Salmon might come to know it.

### 4b. Improving Brueckner's Argument (and Boghossian's)

Brueckner's and Boghossian's arguments failed because they relied on knowers knowing conditionals which they clearly don't know. But their intent was clear enough. Both thought that externalism would show that some knowers know some relevant conditional such that its consequent wasn't known. Were they simply wrong to think so?

To know the answer, it would be useful to understand why such claims might seem plausible. When Brueckner claims that if he knows that he is thinking that some water is dripping, then he knows that he is not thinking that some

twater is dripping, he does so because it seems clearly true that

> [i]f I know that some water is dripping, then I know that I am not a massively mistaken brain in a vat inhabiting a waterless world.[34]

He thinks that (C3) is a second-order analogue to this obviously true conditional; but as we've seen, it isn't. The problem is that the reason the above conditional is true is that knowers have the concept of being a brain in a vat in a waterless world and can distinguish such a concept from the concept water. Knowers don't have a distinct concept of twater when they can't distinguish water from twater, so the second-order analogue to this conditional isn't plausibly knowable.

What gets in the way, then, is lacking the relevant concepts. If one could construct a plausibly knowable conditional which didn't require one to have such concepts, one would have a much better skeptical argument. No doubt, we couldn't appeal to descriptive concepts of twater, because conditionals such as those below are also clearly not knowable:

---

[34] Brueckner (1990), 448.

> If I am thinking that some water is dripping, then I am not thinking that the liquid with chemical composition XYZ is dripping.
>
> If I am thinking that some water is dripping, then I am not thinking that the odorless, tasteless (etc.) liquid on Twin Earth is not dripping.

What Brueckner needs is some way of exporting such descriptions so that they are outside the scope of the 'that' clause. A better second-order analogue, then, is:

> if I know that I am thinking that some water is dripping, then I know that I am not in a waterless world thinking of some substance other than water that it is dripping.

Clearly, anyone who has the concept water also has the concept substance other than water; and clearly it's possible to attribute the belief that some twater is dripping with such de re locutions as the one above. (If one is thinking that some twater is dripping, then one is in a waterless world thinking of some substance other than water that it is dripping.)

Perhaps, then, if we used the following three premises, a better argument would result:

(C1A) If externalism is true, then if I am thinking that some water is dripping, I am not in a waterless world thinking of some other substance that it is dripping.

(C3A) I know introspectively that if I am thinking that some water is dripping, then I am not in a waterless world thinking of some substance other than water that it is dripping.

(C5A) I can't know introspectively that I am not in a waterless world thinking of some substance other than water that it is dripping.

The argument from (C1A)-(C3A)-(C5A)-(C6) to (IT) is clearly superior to Brueckner's argument, and is in the same skeptical spirit. It, too, seems plausibly sound.

Boghossian's argument also benefits from (C3A) and (C5A). Simply substitute (D1A), (D3A) and (D5A) for (D1), (D3) and (D5) and the result is a much stronger argument:

(D1A) If externalism is true, then if I am thinking that water is a liquid, I am not in a waterless world thinking of some other substance that it is a liquid.

(D3A) If I am a switching subject, then the possibility that I am in a waterless world thinking of some substance other than water that it is a liquid is a relevant alternative.

(D5A) I can't know that I am not in a waterless world thinking of some substance other than water that it is a liquid.

Now, we saw that both Brueckner's and Boghossian's arguments depend on some notion of introspective knowledge, and the notion of introspection was nowhere explicated--they assumed that it was clear enough in context. But because

they surely intend to contrast introspective knowledge with empirical knowledge, such knowledge is at the very least non-empirical, in the sense adumbrated in Chapter II: to know introspectively what one is thinking, one must know what one is thinking in a way that doesn't depend on one's knowledge of empirical facts. Once we make this clear, the arguments of Brueckner and Boghossian become identical:

(G1) If externalism is true, then if I am thinking that some water is dripping, then I am not in a waterless world thinking of some other substance that it is dripping.

(G2) If we have privileged access, then I know non-empirically that I am thinking that some water is dripping.

(G3) I know non-empirically that if I am thinking that some water is dripping, then I am not in a waterless world thinking of some other substance that it is dripping.

(G4) I can't know non-empirically that I am not in a waterless world thinking of some other substance that it is dripping.

(G5) If x knows non-empirically both p and that if p then q, x can know q non-empirically.

(G6) Therefore, if externalism is true, then I don't know non-empirically that I am thinking that some water is dripping.

(IT) Therefore, privileged access and externalism are incompatible.

The notion of a relevant alternative drops out as irrelevant, for two reasons: first, because for the sake of argument we are assuming that Closure holds, and second, because if introspective knowledge is independent of the character of one's environment, Boghossian must accept that some restricted version of non-empirical Closure holds--that is, it must hold at least for alternatives involving changes in the character of one's environment. (And because the only cases we're interested in involve such alternatives, we can safely ignore the rest.) Also, if Brueckner and Boghossian rely on such premises as (G1) and (G3), it clearly doesn't matter whether or not individuals have the relevant twearthly concepts: neither premise requires individuals to have such concepts.

Still, there's a problem. (G1) says that externalism implies that thinking about some substance other than water is incompatible with thinking that some water is dripping. One may not know this, and so may <u>not</u> know non-empirically (or even believe) the knowledge claim in (G3). But this is easily remedied: surely it's <u>possible</u> to know the knowledge claim in (G3) non-empirically--someone like Nathan Salmon,

who knows non-empirically that externalism is true, surely does. So we must emend the argument:

(G1A) If externalism is true, then Nathan Salmon knows a priori, hence non-empirically, that if he is thinking that some water is dripping, then he is not in a waterless world thinking of some substance other than water that it is dripping.

(G2A) If we have privileged access, then Nathan Salmon knows non-empirically that he is thinking that some water is dripping.

(G3A) If x knows non-empirically both p and that if p then q, x can know q non-empirically.

(G4A) Therefore, if both externalism and Privileged Access are true, then Nathan Salmon can know non-empirically that he is not in a waterless world thinking of some substance other than water that it is dripping.

(G5A) But Nathan Salmon can't know non-empirically that he is not in a waterless world thinking of some substance other than water that it is dripping.

(IT) Therefore, Privileged Access and externalism are incompatible.

Let's call this argument the Improved Skeptical Argument, or ISA for short.

### 4c. Comparing IMA and ISA

Both IMA and ISA depend on the existence of sophisticated thinkers who are capable of knowing

externalism non-empirically; they also both depend on non-empirical Closure, and on knowledge of conditionals linking what they're thinking to empirical facts they clearly don't know to obtain. So although only ISA is a truly skeptical argument, they have much in common. In fact, the only thing that makes ISA different from IMA is that the consequent of the relevant conditional,

> I am not in a waterless world thinking of some substance other than water that it is dripping,

is unlike standard externalistic content-determiners such as the proposition that I am in an environment containing water. But it is arguably an externalistic content-determiner, in that my thinking that some water is dripping supervenes on this proposition: in any world in which that proposition fails to obtain, I am not thinking that some water is dripping. So ISA, though slightly different in structure from IMA, in fact collapses into it: if we let the consequent of its relevant conditional be labeled 'E', then the two arguments are in all essential respects identical.

Why does this happen? It happens because the 'skeptical' possibilities raised by ISA are entailed by the externalistic consequences of the sort discussed in Chapter

III, and any consequence of an externalistic consequence is itself an externalistic consequence. It should also be clear that these consequences are non-empirically knowable to be consequences of externalistic consequences; so we should expect them to be non-empirically knowable if the externalistic consequences themselves are.

This is a pleasant result: the three arguments for (IT), blemishes removed, turn out to be three modes of presentation of the very same argument. The intuition that externalism and Privileged Access conflict does in fact have a prima facie sound basis, made manifest by IMA/ISA. Because the externalistic content-determiners discussed in ISA are non-empirically knowable consequences of those discussed in IMA, we can treat ISA as a corollary of IMA. So we can treat IMA as the master argument for (IT).

## 5. An interesting corollary

We've noted that ISA is a corollary of IMA, because the allegedly only empirically knowable externalistic facts of the one are non-empirically knowable to be consequences of the allegedly only empirically knowable externalistic facts of the other. In motivating ISA, I've appealed to standard

skeptical considerations and tried to show how they can be adapted in such a way as to apply to privileged access. But it's worth noting that an IMA-style argument can be used to show that radical skepticism of the contemporary, brain-in-a-vatish sort can't be correct. Here's the argument:

> Nathan Salmon knows non-empirically that he's thinking that some water is dripping. He also knows non-empirically that if he's thinking that some water is dripping, then he's not a brain in a vat in a waterless world. So, by non-empirical Closure, Nathan Salmon can know non-empirically that he's not a brain in a vat in a waterless world. If radical skepticism is true, then Nathan Salmon can't even know that he's not a brain in a vat in a waterless world. Therefore, radical skepticism is false.

One might think that this argument bears an uncanny resemblance to Putnam's notorious (1981) anti-skeptical argument. And it does. Inessential features aside, it's precisely the same argument! The reason is simple. If we allow that anything which follows from an externalistic consequence is itself an externalistic consequence, then the proposition that Nathan Salmon isn't a brain in a vat in a waterless world is an externalistic consequence which, by the IMA-style argument above, he can know non-empirically.

A clarification and a caveat are necessary at this point, so as to make things as precise as can be. The

clarification: we have to assume, as Putnam does, that nothing acts on a brain in a vat in such a way as to make it plausible that it has thoughts with normal contents. So if Nathan Salmon is a brain in a vat, nothing must make it plausible that he has thoughts about water, beaches, and sunshine, for example. We can't let an evil scientist program the computer controlling the vat, then, because the scientist's intentions to deceive might somehow determine that the brain has normal contents. So let's stipulate that the computer simply came into existence as it is, so that it's sending signals to the envatted brain in such a way that the brain happens to have experiences which are introspectively indistinguishable from Nathan Salmon's. The caveat: it's plausible that if Nathan Salmon were only recently envatted, his thoughts wouldn't change content. In that situation, he can't appeal to an IMA-style argument to show that he isn't now a brain in a vat in a radically different (e.g., waterless) world. So we ought to weaken the claim known non-empirically: what he can infer non-empirically from externalism and privileged access is that he hasn't always been a brain in a vat in a radically

different world. Even so, to have to constantly spell out the scenario and the claim known would be tiresome. So let me stipulate henceforth that to be a Putnamian brain in a vat is to have always been a brain in a vat of the sort stipulated above.

Now, Putnam himself put the argument differently. He argued on the basis of the semantics of our own utterances that we can't be Putnamian brains in a vat, on the following grounds: first, that our utterances would mean something they don't actually mean were we such brains in a vat;[35] second, that we know that our utterances do mean what they actually mean.[36] But these differences are inessential. Where Putnam relies on the view that we know what our utterances mean in a way that can't be challenged by the radical skeptic, I rely on the view that we know what our thoughts are in a way that can't be so challenged; we both rely on propositional contents being known in the same way. Putnam claims that externalism shows that we are brains in a

---

[35] Also, henceforth I will drop the modifier 'Putnamian'; Putnamian brains in a vat will always be the relevant brains under discussion.

[36] For a thorough analysis and critique of Putnam's argument, see the excellent Brueckner (1986).

vat iff our contents are different from what they actually are; I claim, equivalently, that externalism shows that we aren't brains in a vat iff our contents are as they actually are. Putnam relies on non-empirical Closure, but implicitly; my reliance is explicit. All else is the same.

This is an interesting result, though not an entirely new one. A few others have noticed a relationship between Putnam's anti-skeptical argument and arguments for (IT).[37] But unifying McKinsey's, Brueckner's, and Boghossian's arguments for (IT) has helped us see that we can include Putnam's anti-skeptical argument among the reasons for believing (IT). That's because we can add to the argument above the rider,

> But Nathan Salmon can't know at all, much less know non-empirically, that he's not a brain in a vat,

and thereby conclude that externalism and privileged access are incompatible. Let's call this argument the Anti-Putnamian argument for (IT).

I think this shows the full range of the master argument for (IT), and the full burden of our task. IMA has

---

[37] Brueckner (1990) connects the two, as does Warfield (1995).

two corollaries: ISA and the Anti-Putnamian argument. Any refutation of IMA, then, should also have the resources to block its corollaries.

But enough. We know what we're up against. Let's move on now to some arguments that (IT) can't be correct.

# CHAPTER V

# PREVIOUS SOLUTIONS TO THE PARADOX

## 1. Introduction

In this chapter I consider various solutions which have been offered to the paradox. It's interesting that there are many more solutions to the paradox than there are arguments motivating it. In the beginning of the last chapter, I noted that writers on the paradox tended to take for granted that the paradox is well-motivated. Taking this for granted, though, has its risks. Any true solution to a paradox must explain not only why (at least) one of its component propositions is false, but must also explain why it seems to be true. A paradox is not just any set of inconsistent propositions: each proposition must be plausible on its own. So if one takes for granted that the paradox is well-motivated, one's solution may not take into account the best arguments for each of its component propositions.

As it happens, none of the solutions discussed here is ultimately satisfactory, and only one of them takes the best arguments for (IT) into account. Even so, I think they are all worth discussing, because seeing where they fall short of a true solution may help us discover the way to one.

## 2. Davidson's first solution

Donald Davidson offers two different solutions to the paradox, both of which are contained in his (1987). In this section I examine the first, and examine the second in the next section.

According to Davidson's first solution, the paradox results from taking the metaphor of an object's being before the mind too literally:

> The basic difficulty is simple: if to have a thought is to have an object 'before the mind', and the identity of the object determines what the thought is, then it must always be possible to be mistaken about what one is thinking. For unless one knows everything about the object, there will always be senses in which one does not know what that object is.[1]

For Davidson, propositions are these objects in the case of the attitudes; so the paradox results from thinking about

---

[1] Davidson (1987), 63. Note that page numbers are from its appearance in Cassam (1994).

attitudes towards propositions in a way analogous to the perception of objects.

Why does he think the paradox results from this? Because, in his view, some of the conclusions drawn from the thought experiments of Putnam and Burge seem to indicate that the lack of authority about one's own mental contents derives solely from the claim that contents are determined extracranially--that they 'ain't in the head', as Putnam put the point.

One such conclusion, drawn by Andrew Woodfield (1982), is particularly stark:

> Because the external relation is not determined subjectively, the subject is not authoritative about that. A third person might well be in a better position than the subject to know which object the subject is thinking about, hence be better placed to know which thought it was.[2]

It's worth noting that Woodfield thinks this point so obvious as to be not worth elaborating. But the way he puts the point is significant. It's solely because content-determining relations depend on facts external to thinkers that he concludes that thinkers are not authoritative about

---

[2] Woodfield (1982), viii.

the contents of their own thoughts. The chain of inference is as follows: thinkers are not authoritative about the objects they're thinking about, because these objects aren't mental ('subjective'); the objects thinkers are thinking about determines what thoughts they're thinking; so thinkers aren't authoritative about what thoughts they're thinking.

Woodfield's appeal to the difference between 'subjective' and 'objective' determination of contents does seem to evoke the metaphor of propositions (or their components) being 'before the mind'. His phrasing suggests he thinks that contents are determined 'subjectively' if they are known authoritatively. It's easy to see why if we understand contents to be conscious mental objects of some sort, such as mentalistic modes of presentation: the notion that the objects of consciousness are somehow 'self-revealing' is one of the major differences often alleged to hold between mental and physical objects. If mental contents are not determined by some kind of acquaintance relation between conscious minds and their objects, then someone of a Cartesian turn of mind might well think that they are <u>not</u> known authoritatively.

Davidson thinks this way of motivating the paradox is confused. He alleges that it's correctness depends on the following two untenable assumptions:

> (1) If a thought is identified by a relation to something outside the head, it isn't wholly in the head. (It ain't in the head.)
>
> (2) If a thought isn't wholly in the head, it can't be 'grasped' by the mind in the way required for first-person authority.[3]

What, according to Davidson, is wrong with (1) and (2)? The falsity of (1) seems to him transparent:

> To suppose [(1)] would be as bad as to argue that, because my being sunburned presupposes the existence of the sun, my sunburn isn't a condition of my skin. My sunburned skin may be indistinguishable from someone else's skin that achieved its burn by other means ... yet one of us is sunburned and the other not.[4]

The reasoning implicit in Davidson's example is as follows. Sunburn is obviously a condition of my skin, yet <u>it</u> is identified by relation to something outside the skin, namely the sun. By parity of reasoning, then, thoughts should not be thought to not be in the head because they're identified by relation to something outside the head.

---

[3] Davidson (1987), 58.

[4] Ibid., 58.

The problem with Davidson's reasoning here is that no one denies that sunburn is a condition of the skin, or that thoughts are in the head. In a very important sense, sunburn has to be in the skin, and thoughts have to be in the head: for sunburns and thoughts to have the causal powers they actually have, they must be physically realized. But all this means is that sunburn and thought realizations are in the skin and the head, respectively. Of course, neither does anyone think that the property of being sunburned or the property of being a thought are in the skin or the head (or, indeed, anywhere at all). And it's the properties which are determined by relations to extradermal or extracranial factors, not their physical realizations.

Now, to deny that properties are literally in the head is to make a trivial claim. Putnam wasn't making a trivial claim; so he must have been denying something else.[5] This means, quite simply, that his remark that meanings 'ain't in the head' should be taken metaphorically, as he no doubt

---

[5] That Davidson, himself an externalist, makes this mistake suggests that this way of understanding Externalism (shared by McKinsey) is rooted in the sloppy ways in which the thesis is usually formulated.

intended it to be. What he denied was that the reference of singular and natural-kind terms is determined by descriptions (or similar things, such as conceptual roles, stereotypes, and "gestalts") speakers may associate with them. Because this entails that the propositions two speakers express in uttering the same sentence may differ even though they associate all and only the same descriptions (conceptual roles, stereotypes, "gestalts") with that sentence's relevant terms, it follows that the thoughts they are expressing with that sentence may also differ. That meanings 'ain't in the head' means that meanings aren't determined only by those features of meaning accessible to speakers, but also by factors which, because they are essentially extracranial, may be inaccessible to speakers. Understood in this (correct) way, (1) is not refuted by Davidson's sunburn example.

So much for (1). What motivates Davidson's rejection of (2)? He elaborates on the motivation for (2) as he understands it:

> True, my sunburn, though describable as such only in relation to the sun, is identical with a condition of my skin.... Still, if, as a scientist skilled in all the physical sciences, I have access only to my skin, and am

> denied knowledge of the history of its condition, then by hypothesis there is no way for me to tell that I am sunburned.... The difference between referring to and thinking of water and referring to and thinking of twater is like the difference between being sunburned and one's skin being in exactly the same condition through another cause. The semantic difference lies in the outside world, beyond the reach of subjective or sublunar knowledge.[6]

The idea behind this argument seems to be the Cartesian idea that we have authoritative knowledge of mental content if this content is 'subjective', but that if it's instead part of the 'objective' outside world, then we have no privileged access to it. And Davidson rightly criticizes this idea:

> This analogy, between the limited view of the skin doctor and the tunnel vision of the mind's eye, is fundamentally flawed. It depends for its appeal on a faulty picture of the mind.[7]

This picture is that of mental objects being 'before the mind', so that we come to know about physical objects through the prism of their mental correlates.[8] If Externalism entails that propositions are not identifiable

---

[6] Ibid., 60.

[7] Ibid.

[8] Of course, a view superficially like this one--viz., that to know about things presupposes mentally representing them--is obviously true. But the obviously true view has no obviously consequent epistemology, and clearly doesn't entail sense-datum-style theories.

with mental objects by virtue of their dependence on extracranial factors, then if one holds this picture one will think of propositions as being like physical objects in a crucial respect: we can know about them only via their mental correlates, in this case 'narrow' psychological states, only to which we have authoritative access.

This long-discredited view is also known under other names: the Myth of the Given, the Myth of the Museum, the Cartesian Theater. We don't need to go on at length about why it is untenable. If the paradox depends on this view of the mind, then we clearly have no good reason for thinking there is a paradox. But why should we think the paradox does depend on this view?

To be sure, some arguments for (IT) may depend on a manifestly Cartesian picture of the mind. The line of reasoning Davidson characterizes and rebuts shows that these arguments may well seem tempting. But Davidson nowhere addresses either IMA or its corollaries, which offer the best reasons for believing (IT).

Perhaps Davidson would claim that IMA in some implicit way depends on the Cartesian picture of the mind. But if

he's to offer a solution to the paradox, he must explain how IMA does depend on this picture, especially since it seems not to. Neither IMA nor its corollaries explicitly claim that propositions must be 'before the mind' if they are to be known authoritatively; all they explicitly depend on (apart from externalism itself and its obvious consequences) are non-empirical Closure and the non-empiricality of privileged access. Davidson wouldn't challenge the non-empiricality of privileged access (although he describes the phenomenon differently, as "First Person Authority"--see the next note). And it's hard to see what non-empirical Closure has to do with the Cartesian picture of the mind.

So Davidson's first solution is, on its face, not so much false or implausible as irrelevant. It rebuts clearly inferior arguments for the paradox, ignoring the best.

### 3. Davidson's second solution: the Davidson-Heil Thesis

To be fair, Davidson regards his first solution to the paradox as really only half of a solution. His intention is to explain privileged access, and he thinks that externalism, far from being a thesis which makes privileged access mysterious, actually helps explain it:

> If we can bring ourselves to give up this picture [that of the Cartesian theater of the mind], first-person authority will no longer be seen as a problem; indeed, it will turn out that first-person authority is dependent on, and explained by, the social and public factors that were supposed to undermine that authority.[9]

How does externalism help explain privileged access, according to Davidson? By virtue of its recognition that

> what a person's words mean depends in the most basic cases on the kinds of objects and events that have caused the person to hold the words to be applicable; similarly for what the person's thoughts are about.[10]

The Putnam-Burge thought experiments have presumably shown just that. What's more, they have shown that a thinker's knowledge of the determinants of her thought and utterance contents is not required for her to have the thoughts or meaningfully utter the sentences whose contents they determine. Indeed, typically she is not:

> The agent herself ... is not in a position to wonder whether she is generally using her own words to apply to the right objects and events, since whatever she

---

[9] Davidson (1987), 60-61. Davidson uses the term 'first-person authority' to denote non-empirical knowledge of one's own mental states, because he dislikes the traditional connotations of 'privileged access'. See Davidson (1984) and my discussion of it in Chapter VII.

[10] Ibid., 64.

> regularly does apply them to gives her words the meaning they have and her thoughts the contents they have.[11]

Davidson thinks this intimate link between a thinker's usage and her social and environmental circumstances plays a key role in explaining privileged access; but more is clearly needed. Why should the fact that whatever a thinker applies her words to gives those words the meaning they have entail or even suggest that a thinker has special authority about the contents of her thoughts? Davidson claims that the need for charity explains the entailment:

> Unless there is a presumption that the speaker knows what she means, i.e. is getting her own language right, there would be nothing for an interpreter to interpret. To put the matter another way, nothing could could as someone regularly misapplying her own words.[12]

This explanation is not very plausible. The need for charity in the interpretation of speech has no obvious connection with any special authority about the contents of one's thoughts.

It's worth noting, though, that a solution to the paradox which rejects (IT) may not call for an explanation of privileged access; all it definitely requires is an

---

[11] Ibid.

[12] Ibid.

explanation of why privileged access and externalism don't conflict. (A solution may call for an explanation of privileged access in that it depends on features of privileged access which (perhaps only partly) explain it.) As I noted in the introduction to this chapter, this really amounts to an explanation of two things: (i) why the best arguments for (IT) fail; (ii) why (IT) can seem to be true. So even if Davidson's explanation of privileged access is no good, one may take aspects of it in attempt to explain (i) and (ii). This is what John Heil does, in his (1988).

Heil's strategy is straightforward. Twins differ not only in their thoughts about the world, but also in those thoughts of theirs which are about those thoughts. Believing that one is thinking that water is a liquid is not identical to believing that one is thinking that twater is a liquid, for reasons identical to the reasons that the first-order thoughts are not identical. So the same factors determine that both believing that water is a liquid and believing that one is thinking that water is a liquid involve the concept water.

Once we recognize this, we must also recognize that if having thoughts about water doesn't require knowing that the factors which determine such thoughts obtain, then having thoughts about thoughts about water doesn't require knowing that those factors obtain. This epistemic point, according to Heil, is a significant one, because he thinks the paradox depends on the assumption that we must have access to the conditions which determine that beliefs about one's own thoughts are in fact about those thoughts:

> It might seem at first blush that access to the contents of [my] first-order states ... would necessitate my somehow coming to recognize the obtaining of states of affairs ... responsible for first-order content.[13]

Let's call the conjunction of the determination point and the epistemic point the Davidson-Heil Thesis (DH).[14] Then Heil's argument against (IT) is twofold: first, that it depends on the view that our access to our mental contents requiring knowledge that their determining conditions obtain; second, that DH repudiates this view.

---

[13] Heil (1988), 244.

[14] The Davidson-Heil Thesis was so dubbed by Schiffer (lectures).

Two problems with Heil's argument arise at once. First, it's not at all clear how (IT) depends on this view, and Heil nowhere explains why privileged access, the non-empirical knowledge of one's mental states, should even prima facie require knowing that the determining conditions of those states obtain. Indeed, the requirement that one know that the determining conditions of a mental state obtain seems to rule out the possibility of one's access to that state's contents being privileged. Second, it's not immediately clear how DH repuduates this view. If by 'access' Heil means privileged access, then DH doesn't seem to have anything to do with a requirement (however implausible) that non-empirical knowledge of a mental state involve knowing that the determining conditions of that state's content obtain.

The reasons for Heil's second claim become clear, however, if we suppose that what Heil means by having 'access' to a thought's content just is having a thought about that thought. Sometimes he writes as if this is what access amounts to. For example, in the passage above, he is using the idea that access to one's thoughts requires

knowing that their determining conditions obtain in order to motivate a negative answer to the following question:

> Is it plausible to suppose that [the content of my introspective thought] includes that of ... my first-order mental state?[15]

This is a very strong form of skepticism, one directed not against privileged access, but instead against the view that one has beliefs about one's own mental states. Indeed, in one place he characterizes the argument for (IT) as being that of "a nastier sceptic, one who questions the presumption that we think what we think we think."[16]

But he also characterizes access as a kind of privileged access, and then he is writing about knowledge of one's own mental states. Immediately upon asking the question above, he asks another:

> And even if this is so [that my introspective state includes the content of it's first-order object], is there any reason to think either that [my first-order thought's] content, whatever it is, could be accurately preserved in my introspective thought, ... or that my access to the content of [my first-order thought] could be in any sense epistemically direct?[17]

---

[15] Ibid.

[16] Ibid., 245

[17] Ibid.

If he's concerned about the accuracy or epistemic directness (i.e., non-inferentiality) of introspective thoughts, then he is concerned not about whether having thoughts about thoughts requires knowing that the latter's determination conditions obtain, but about whether knowing that one is having a thought requires knowing that that thought's determination conditions obtain.

Unfortunately, Heil never distinguishes between these two points. This allows him to claim that DH shows that we have access to our mental states in both senses of 'access': that of having thoughts about mental states and that of having non-empirical knowledge of those states. But DH shows that access to our own mental states doesn't require knowing that their determining conditions obtain only in the first sense of 'access'.

What's worse, DH in no way addresses any of the premises of either IMA or its corollaries. At best, DH demonstrates that externalism is compatible with our having non-empirical beliefs about our own mental states; but such beliefs are not denied by the best arguments for the

paradox.[18] So we cannot accept DH as in any sense a solution to the paradox, because it explains neither why (IT) is false nor why (IT) seems to be true.

Despite this, DH does show something interesting about the relationship between thoughts and beliefs about them: that whatever factors determine that a thought of the form x is F involves the concept F also determine that a thought of the form S is thinking that x is F involves F. This raises a tantalizing possibility. If one believes that one is having a thought that ... F ..., then (granting the assumption below), it's simply not possible for one to do so while (i) not having any thought that ... F ... and (ii) instead having a thought that ... F' ..., where F and F' are not distinct concepts.

---

[18] One might generate arguments analogous to IMA using belief instead of knowledge. The upshot of these arguments would be that because no one can believe non-empirically that externalistic consequences of one's thoughts obtain, privileged access and externalism are incompatible.

Unfortunately, it's quite plausible that one can believe externalistic consequences in a way that doesn't depend on having empirical evidence--this plausibility just is what motivates DH. Therefore, the simplest response to such arguments for (IT) is to point this out, as Heil does.

This possibility depends on the plausible assumption that we can't have higher-order beliefs of the sort had in having privileged access in such a way that those beliefs involve one set of concepts and the conscious mental states those beliefs are about involve indistinct concepts.

The assumption is plausible because the beliefs by virtue of which we have privileged access are states by virtue of which we're conscious of our conscious mental states, and so they come into being more or less contemporaneously with the states they're about. And it's virtually impossible to imagine how someone could more or less contemporaneously come to be in two mental states involving different concepts which aren't distinct from one another.

Given this assumption, such skeptical arguments as ISA can seem difficult to take seriously. If we can't be wrong about those mental states to which we have privileged access in the way ISA requires, then it's hard to see why we should believe it's sound. And because ISA is a corollary to IMA, it may be hard to see why IMA should be sound as well. I think that the plausibility of this assumption and the

considerations it raises are what make DH so compelling as a solution to the paradox. Unfortunately, the considerations it raises simply reinforce the sense of paradox; they do not resolve it.

We have rejected DH as a stand-alone solution to the paradox, because by itself it simply doesn't address it. But we have not rejected every solution which has DH as a component. It may be possible to extend DH in such a way as to give force to its claim that the Putnamian thought experiment doesn't in any way undermine privileged access. Indeed, there are two other solutions employing DH which I will address in this chapter: one due to Falvey and Owens (1994), which conjoins DH with a denial of Closure; and the contextualist solution, which conjoins DH with a contextualist explanation of IMA and its corollaries are unsound. So we're not by any means through with DH.

## 4. Burge's solution

The solution of Tyler Burge (1988) shares some features with DH, in that it, too, appeals to the fact that the same factors determine relevant components of the contents of both first-order and second-order thoughts. But even though

Burge's solution, too, fails, the fundamental idea motivating Burge is interesting and important, even if it doesn't by itself lead to a solution to the paradox. Let's see this idea first, then see how Burge uses it to solve the paradox.

Burge's idea is to take Descartes's Cogito as the paradigm case of privileged access, and explain why the Cogito and Cogito-like judgments have the authority they have independently of Externalism. Burge's examples of such judgments include: that I am now thinking; that I think (with this very thought) that writing requires concentration; that I judge that water is more common than mercury. Such judgments, according to Burge, are "self-verifying in an obvious way: making these judgements itself makes them true."[19] He calls these judgments basic self-knowledge.

But basic self-knowledge is not merely self-verifying; it is so because it has the peculiar feature of also being self-referential:

---

[19] Burge (1988), 66. Note that page numbers are from its appearance in Cassam (1994).

> In the case of perceptual knowledge, one's perception can be mistaken because some counterfeit has been substituted. It is this possibility which tempts one into the (mistaken) view that, to have perceptual knowledge, one must first know something that rules out the possibility of counterfeit. But in the cases of cogito-like self-verifying judgements there is no possibility of counterfeit.... Basic self-knowledge is self-referential in a way that ensures that the object of reference just is the thought being thought.[20]

How is basic self-knowledge self-referential? Burge's explanation is vague, but he clearly thinks that basic self-knowledge exhibits the character of the Cogito; so it may be helpful to see why one might think that the Cogito is self-referential. Consider the judgment:

I am now thinking.

When I make that judgment, I am referring to an act of thinking, indeed, to the very act of thinking which constitutes that judgment. We might render this fact more perspicuous by making it explicit in the judgment itself:

I am, with this very thought, now thinking.

Now we may ask how we know this judgment to be true. Well, we know it must be true, because its very occurrence guarantees its truth--it's logically impossible to make such

---

[20] Burge (1988), 74-5.

a judgment and be mistaken. Every such judgment is a true one, partly because in making it one is referring to the very judgment one is making. This account applies to such judgments as that I judge that water is more common than mercury in the following way: in the act of judging that one is judging that water is more common than mercury, one is judging that water is more common than mercury, because the sense of the second-order use of 'judge' is performative--by 'I judge' Burge means 'I hereby judge'.[21]

What does this mean? To see, let's consider an obvious performative notion: promising. When one is promising by uttering the sentence

I hereby promise to mow your lawn,

---

[21] This is not the only possible interpretation of Burge's notion of self-reference. One might understand Burge to be asserting DH in a more roundabout way. One could claim that the contents higher-order thoughts "contain" the contents of lower-order thoughts as constituents, so that the thinking of a higher-order thought is as were an indirect thinking of a lower-order one. (I owe this point to Schiffer.)

Still, I prefer the performative interpretation, mainly because it affords a clear explanation of how self-reference leads to self-verification; DH, even supplemented by a vague notion of "containment", offers no comparable explanation.

one is not merely promising, but also referring to the very act of promising one is making. By analogy, in performatively judging that one is thinking a thought, one is not merely judging that one is thinking that thought (by referring to that thought in the act of thinking it), but also thinking the very thought one judges oneself to be thinking. So we must understand basic self-knowledge to be constituted by such judgments as:

> I am (with this very judgment) now judging that water is more common than mercury;
>
> I am (with this very judgment) now thinking that writing requires concentration;

where these higher-order judgments are themselves instances of lower-order ones, by virtue of the performative notion of judgment.

If privileged access amounts to basic self-knowledge, Burge thinks, then there can be no conflict between externalism and privileged access. That's because he takes the problem of reconciling externalism and privileged access to be one of explaining how, given externalism, we can know what mental states we're in without having knowledge of what

determines their contents. That is, for Burge, (IT) is motivated by an argument of the following sort:

> If externalism is true, then we lack privileged access to the determinants of mental content. But if we lack privileged access to the determinants of mental content, we lack privileged access to mental contents. Therefore, if externalism is true, we lack privileged access to mental contents.

This argument is reminiscent of Brueckner's skeptical one, which we discussed and rejected in the previous chapter. But if it were the best motivation for (IT), then Burge's characterization of basic self-knowledge would help solve the paradox, if privileged access is basic self-knowledge.[22]

Unfortunately, IMA and its corollaries are the best motivations for (IT), and privileged access is not basic self-knowledge. Let me explain each point in turn.

IMA and its corollaries assert that privileged access is non-empirical knowledge. Clearly, Burge intends that basic self-knowledge be non-empirical as well. But he doesn't acknowledge the possibility that one can also know

---

[22] I wouldn't go so far as to say that it solves the paradox outright, because it's not clear how Burge would explain (IT)'s appeal. If having privileged access is simply a matter of making self-referential and self-verifying judgments, why would any sensible person think it conflicts with externalism?

externalism and its logical consequences non-empirically, and thereby come to know *prima facie* empirical facts non-empirically. Perhaps he would deny the non-empiricality of externalism (Chapter IV, note 34). But because he is mute on the subject, his solution is, at best, incomplete.

What's more, Burge's notion of basic self-knowledge can't account for privileged access. For privileged access to be basic self-knowledge, it would have to be that *every* mental state to which we have privileged access is a self-referential, self-verifying judgment. Clearly, this is not so. We can see why by considering the case of sensations, but it's also clear that even garden-variety intentional states are dubious candidates for basic self-knowledge.

We have privileged access to our sensations, but not only is it not even clear that sensations *have* intentional content, it's hard to see how, say, a pain can be identical to a *judgment* that one is in pain, much less one that is self-referential and self-verifying. Of course, one might reflect on one's pain and conclude that, with that very pain, one is in pain. But the pain itself is not identical to the conclusion that one is in pain, so *a fortiori* not

identical to a self-referential, self-verifying one. The same holds (mutatis mutandis) for other sensations.

Intentional states also rarely exhibit this character, but we seem to have privileged access to many of them as well. Boghossian points out that we have privileged access to dispositional states such as beliefs and desires, but that judgments that one believes or desires such and such can't be self-referential or self-verifying.[23] (Judging

---

[23] Boghossian (1989), 21. Interestingly, Boghossian goes on to argue (22-3) that even basic self-knowledge is incompatible with externalism. His argument is to the effect that because, in recalling what was once an item of basic self-knowledge concerning water, one won't know whether the item was a thought about water or a thought about twater; so either one forgot, or one never knew. But because it's not plausible to think that one forgot, one never knew.

Boghossian overlooks one important point, however. It must be established that because one doesn't know whether one's thought is about water or twater, one's judgment that one is thinking that ... water ... can't be knowledge. Burge (1988, 78) explicitly denies this in the case of basic self-knowledge:

> One should not assimilate "knowing what one's thoughts are" in the sense of basic self-knowledge to "knowing what one's thoughts are" in the sense of being able to explicate them correctly--being able to delineate their constitutive relations to other thoughts.

(See Falvey and Owens 1994 for a more extended discussion of this distinction, which I examine in Section 6 below.) Perhaps Burge is wrong to deny this, but we must keep in mind that "basic self-knowledge" is a term he defined; so how would Boghossian show that he is wrong?

that one has a belief or desire isn't itself an act of believing or desiring, simply because there are no acts involved in being in dispositional states.)

To be fair to Burge, he acknowledges that basic self-knowledge doesn't account for all the phenomena which constitute privileged access:

> Basic self-knowledge is at most an illuminating paradigm for understanding a significant range of phenomena that count as self-knowledge. Thus, the whole discussion has been carried out under a major simplifying assumption.[24]

But there's an important question to answer here. We need to know just how significant a phenomenon basic self-knowledge is if we're to know whether it should serve as a paradigm for privileged access generally. If the mental states which constitute privileged access quite often aren't self-referential or self-verifying, then it would seem that the features which make basic self-knowledge what it is don't apply to privileged access as a whole. This implies that whatever gives us special, non-empirical authority over the contents of our own mental states, it's not (in general) self-reference or self-verification. I've offered reason

---

[24] Burge (1988), 79n.

for doubting that basic self-knowledge is a very significant phenomenon. So whether or not Externalism is compatible with judgments about our own thoughts being self-referential and self-verifying, Burge has in no way established that it is compatible with privileged access as a whole.

To his credit, Burge acknowledges this, too. (Although I suspect he thinks basic self-knowledge more significant a phenomenon than I do.) Nevertheless, he doubts that even when we consider privileged access as a whole, it will turn out to be incompatible with externalism:

> I think ... that reflection on the way errors can occur in such cases [i.e., cases in which judgments are not self-verifying or immune to error] gives not the slightest encouragment to the view that anti-individualism (as regards either the physical or social environments) is a threat to the authority of our knowledge of the contents of our thoughts.[25]

I wholeheartedly agree. Still, what we agree on requires substantial argumentation, argumentation we have yet to see.

## 5. Falvey and Owens's first solution

Falvey and Owens (1994) offer, in my view, one of the most serious and comprehensive attempts to solve the

---

[25] Burge (1988), 79n.

paradox. They take themselves to be offering only one solution, but in the same paper they offer a distinct solution to what I call the Anti-Putnamian Paradox, which, as we'll see shortly, is really an instance of the paradox we're discussing here. The solution they take themselves to be offering to the paradox involves denying Closure; neither IMA nor its corollaries appeal to Closure, strictly speaking, so this solution isn't fully adequate. Even so, it's interesting and important, so I will discuss it in the next section.

The other solution they offer involves a different strategy: they deny that one can know non-empirically that believing that one is thinking that ... water ... entails E, where E is some externalistic content-determining proposition involving the concept water. This solution directly addresses a premise crucial to IMA, so it is deserves serious examination, which I begin now.

First, let's take a look at the Anti-Putnamian argument for (IT), stripped to its essentials:

(A1) I can't know, so a fortiori can't know non-empirically, that I am not a brain in a vat.

(A2) Externalism is true.

(A3) If I have privileged access and externalism is true, then I can know non-empirically that I am not a brain in a vat.

Central to the motivation for (A3) is the view that we have privileged access to our own mental contents. A good way of stating the argument for (A3) is as follows:

> If externalism is true, then I can know non-empirically that if I am a brain in a vat, I could not now be thinking the thought that I am not a brain in a vat. But I know non-empirically that I am thinking the thought that I am not a brain in a vat. Therefore, if externalism is true, then I know non-empirically that I am not a brain in a vat.[26]

Falvey and Owens discuss other ways of stating the argument, but settle on this one as best.[27]

---

[26] This version of the motivation for (A3) is very close to the one defended by Michael Williams (1984), Thomas Tymoczko (1989), and Marian David (1991). Mine differs from the other in my making explicit the need for non-empirical knowledge of one's thoughts and of externalistic content-determining conditionals.

One should also notice that the argument contains a suppressed premise: non-empirical Closure.

[27] Anthony Brueckner, who favors a metalinguistic argument for (A3) in his (1986), doesn't agree that it is best: see Brueckner (1994) for discussion. I think it is, but only because it is simplest. Metalinguistic arguments for (A3) rest on non-empirical knowledge of the meanings of one's own utterances, but having privileged access to the meanings of one's own utterances, in my view, isn't essentially different from having privileged access to the contents of one's thoughts.

As we saw in Chapter IV, that I am not a brain in a vat is arguably an externalistic content-determiner of every thought about the external world, in that there is no possible world in which (i) my brain is physically as it actually is, (ii) I am thinking the external-world thoughts I am actually thinking, and (iii) I am a brain in a vat. That is, my external-world thoughts supervene on the truth of the proposition that I am not a brain in a vat, just as my water thoughts supervene on the truth of (let's say) the proposition that there is water in my environment. Therefore, by non-empirical Closure, if I know non-empirically that I am thinking that I am not a brain in a vat, then I know non-empirically that I am not a brain in a vat thinking of myself that I am such and such (whatever being such and such amounts to in an envatted environment).

Because the Anti-Putnamian argument for (IT) is a corollary to IMA, the solution to the Anti-Putnamian Paradox that Falvey and Owens offer, if it works, should also solve the paradox of externalism and privileged access. So what is their solution?

They break the argument for (A3) down into its components:

(A) I know (non-empirically) that I am now entertaining the thought that I am not a brain in a vat.

(B) I know (non-empirically) that if I were a brain in a vat, then I could not entertain that thought.

(C) Therefore, I know (non-empirically) that I'm not a brain in a vat.[28]

Clearly, (A)-(C) is valid (assuming non-empirical Closure). Therefore, either (A) or (B) must be rejected. Falvey and Owens have no objection to (A); their problem is with (B).

I argued that (B)-style conditionals can be non-empirically known in the previous chapter; but was my argument sound? Falvey and Owens think not. They argue that the reasoning justifying the claim that the 'water' thoughts of Earthlings are about water (and that the 'water' thoughts of Twearthlings are about twater) makes essential reference to empirical facts about water (and twater). Concerning Oscar and his twin Twin-Oscar, who live on Earth and Twin Earth respectively, they write:

---

[28] Falvey and Owens (1994, 32) use slightly different sentences, but their arguments apply equally to these.

> If I am ignorant of the chemical composition of water, then I will not be in a position to form [on the basis of describing the difference between Earth and Twin Earth in terms of the difference between $H_2O$ and XYZ] any judgment concerning the reference of Twin-Oscar's words. My ignorance of the fact that water is not XYZ prevents me from recognizing any difference between Earth and Twin Earth. Consequently, for all I know, Twin-Oscar's word 'water' does refer to water. It certainly does if water is XYZ, and this may be the case, for all I know. Thus, the supposition that the twins' thoughts differ in content clearly rests on the empirical premise that water is not XYZ.[29]

If Falvey and Owens are correct, then I can't know non-empirically that if I'm thinking that water is a liquid, then my twin is not thinking that water is a liquid. The reason is that my knowledge of this conditional depends on knowing the following empirical premises: (i) that water is H2O, and (ii) that water is not XYZ.

But is it really true that my knowledge of this conditional depends on knowing those premises? Because I do know them, it may be hard for me to determine whether or not other knowledge I have depends on this knowledge; so the simplest thing to do in this case is to change the example. Let's consider, not water, but rabbits. I know very little about rabbits--more to the point, I don't know enough about

---

[29] Ibid., 134.

rabbits to say what makes them rabbits as opposed to dogs, or cats, or guinea pigs. (I presume here that what makes rabbits rabbits are facts about their genetic makeup.)

Given this, let's suppose that my twin is in an environment in which there are animals which look and behave exactly as rabbits do (as far as my knowledge of rabbits is concerned), but which happen not to be rabbits. It doesn't matter what the reason is--indeed, let's just stipulate it. It follows from Externalism that if my twin has thoughts about these creatures, then his thoughts are not about rabbits. His 'rabbit' thoughts are about distinct creatures, whatever they may be about (apart from rabbits, of course). My 'rabbit' thoughts, by contrast, are about rabbits. Therefore, my having rabbit thoughts entails some externalistic fact about rabbits--say, that there are rabbits in my environment.

Have I appealed to any empirical premise in this reasoning? In particular, have I appealed to any empirical premise akin to (i) and (ii) above? It seems clear that I have not. The crucial premise I appealed to was simply that rabbits are not rabbitesque non-rabbits, and it's analytic

that rabbits are not rabbitesque non-rabbits. (In fact, it's arguably a logical truth, if one wishes to quibble about analyticity. In either case, one needn't investigate the world to know it.)

Perhaps Falvey and Owens would reply that I can't just stipulate that my twin is in an environment containing rabbitesque non-rabbits; how do I know what's in my twin's environment? For all I know about rabbits, they may object, anything which looks and behaves as rabbits do just is a rabbit.

This reply isn't very persuasive. What Falvey and Owens must claim is that for all I know, there can't be rabbitesque non-rabbits. But what gives them such insight into what I know? To know that there can be rabbitesque non-rabbits, I think, is not to know very much. In fact, if the sense of 'can' here is epistemic, to know this is to know that, for all anyone can know, there are rabbitesque non-rabbits.[30] And this seems clearly true.

---

[30] That is, all I need to know is that the property of being a rabbitesque non-rabbit isn't necessarily uninstantiated. Who, I wonder, would deny this?

The analogy with water is clear. The sole purpose of introducing the term 'XYZ' is to introduce a (fictional) substance which, by stipulation, is a substance distinct from water. Particular facts about water aren't relevant here. What is relevant is that there can be a wateresque substance distinct from water; and knowledge of this possibility doesn't seem to rest on knowledge of any empirical facts whatsoever. (Unless Falvey and Owens wish to claim, implausibly, that to know that a wateresque substance distinct from water is possible requires one to know water's chemical microstructure.)

So this solution to the paradox, while it has the merit of actually addressing a corollary to the best argument for (IT), is in the end unpersuasive. One thing it does, though, is help us distinguish between what knowledge of externalistic conditionals can involve and what it must involve--that is, it can but need not involve knowledge of empirical facts.

Before moving on, let me briefly recap. The four solutions discussed thus far each fail to solve the paradox. Three of these fail for the same, obvious reason: none

address the best reasons for believing that externalism and privileged access are incompatible. One solution--Burge's--has the additional failing of offering an implausible positive account of privileged access. The last solution does address the best reasons for believing that externalism and privileged access are incompatible, but it isn't very convincing.

The next (and final) two solutions I discuss each involve in one way or another the Closure principle upon which IMA and its corollaries rest. Each of these solutions is interesting, because each highlights important aspects of privileged access, aspects which may well figure in a true solution to the paradox.

## 6. Falvey and Owens's second solution

With their second solution, Falvey and Owens take the best argument for the paradox to be Brueckner's, an argument we rejected in the previous chapter. But we may take them to be arguing against IMA and its corollaries, for we can easily adapt their solution to fit them.

According to Falvey and Owens, the key mistake made by skeptics of privileged access is to conflate different

notions of knowledge of content. There is introspective knowledge of content, which is a thinker's knowing "the contents of his occurrent thoughts and beliefs authoritatively and directly (that is, without relying on inferences from observation of his environment)."[31] This is (more or less) the notion of privileged access we have been working with. But there is also introspective knowledge of comparative content, which is a thinker's knowing "authoritatively and directly" of any two thoughts "whether or not they have the same content."[32] They admit that externalism is incompatible with the latter, but claim that there are good reasons for thinking we lack introspective knowledge of comparative content anyway. Their main argument, however, is for the claim that externalism is compatible with introspective knowledge of content. Let's briefly see why they think we lack introspective knowledge of comparative content before examining their main argument.

Falvey and Owens think we lack introspective knowledge of comparative content because they think that principled

---

[31] Falvey and Owens (1994), 109.

[32] Ibid., 109-10.

disagreement about whether two thoughts have the same content shows we cannot rely on introspection alone to decide the issue. They cite the Mates-puzzle example as an illustration:

> (1) Nobody doubts that whoever believes that Mary is a physician believes Mary is a physician.
>
> (2) Nobody doubts that whoever believes that Mary is a physician believes Mary is a doctor.

They claim that because Benson Mates believes that (1) and (2) express different thoughts, while Alonzo Church believes that they express the same thoughts, mere introspection won't decide the issue.[33]

> To resolve the dispute between Mates and Church one does not need better inner eyes; one needs additional information about the world we live in, the nature of our linguistic practice, the semantic theories that best represent that practice, and so on. While much of this information is logico-philosophical in character, it is not plausible that it can be acquired independently of a serious empirical investigation into linguistic practice.[34]

---

[33] See Mates (1952) and Church (1954).

[34] Falvey and Owens (1994), 113.

So if Falvey and Owens are correct, our knowledge of comparative content is not always purely introspective, whether or not externalism is true.[35]

I think the issue Falvey and Owens raise is complicated, but that they're basically right. Knowledge of comparative content is not always purely introspective. But we must ask: If we lack introspective knowledge of comparative content because knowing whether two thoughts have the same content can require empirico-linguistic knowledge, why don't we lack introspective knowledge of content because knowing whether one's thought has such and such a content also can require empirico-linguistic knowledge? The answer seems simple enough: that one doesn't know that two thoughts have the same content just doesn't

---

[35] In fact, this is easier to see when we consider, not belief sentences, but mathematical ones. (Externalists and individualists notoriously give different accounts of belief sentences.) For example, a much-debated topic is whether such sentences as 'Triangles are trilateral' and 'Triangles are triangular' have the same content, given that each sentence is necessarily true and that 'triangular' and 'trilateral' necessarily have the same extension. It's clear that one's knowledge of whether such sentences have the same content won't rest on whether externalism is true; so introspective knowledge of comparative content is independent of the status of externalism.

imply that one doesn't know when one is thinking one thought or the other. An analogy from perception helps us understand why. Suppose I see a bird at time $t_1$, and then I see a bird at $t_2$. I don't know whether my seeings are of the same bird, but that doesn't imply that I didn't know what I saw at either $t_1$ or $t_2$.

They go on to claim, though, that the mistake one would make in maintaining (IT) is to think that the kinds of introspective knowledge are the same; and this just doesn't seem true. If I have introspective knowledge of content but lack introspective knowledge of comparative content, then I can know that I'm thinking p without knowing whether my thinking p is identical to my thinking q. But if I don't know this, I won't know that <u>if I'm thinking p, then I'm not thinking q</u>, which is a conditional crucial to the skeptical argument for (IT) they take as their target. So if I lack introspective knowledge of comparative content, I won't know the relevant conditionals upon which skeptical arguments rest, which would mean that a Brueckner-style skeptic who

relied on my lacking introspective knowledge of comparative content to motivate (IT) would undermine her own argument.[36]

Therefore, I can't endorse Falvey and Owens's diagnosis of the mistake made in endorsing (IT). (If, in fact, it is a mistake.) Luckily, this doesn't compromise their own solution to the paradox, which fails for different but illuminating reasons.

Falvey and Owens's solution to the paradox involves explaining how one can know non-empirically what one is thinking without knowing its externalistic consequences. This explanation requires denying non-empirical Closure, of course, but denying non-empirical Closure requires one to explain why privileged access is non-empirical knowledge even though non-empirical Closure doesn't hold--that is, explaining how it's possible to know something non-empirically even though one doesn't know (non-empirically) its (non-empirically) known consequences.

---

[36] To be fair to Falvey and Owens, they are not the only ones who've failed to notice this. Brueckner (1990, 1994), Boghossian (1989), Warfield (1992) and Ludlow (1995) have also overlooked the point. See Chapter IV.

They do so by appeal to a version of the Relevant Alternatives theory of knowledge (henceforth, "RA"), according to which one need rule out only the known relevant alternatives to what one knows to count as knowing.[37] In particular, one needn't know that one is not thinking of some substance distinct from water that it is dripping in order to count as knowing that one is thinking that some water is dripping, even though one knows that thinking the latter rules out the possibility that one is thinking the former. Now, the possibility that one is thinking of some substance distinct from water that it is dripping may well be a relevant alternative, depending on the context.[38] So RA can't by itself explain why we can know that we're thinking that some water is dripping but not that we're not thinking of some substance distinct from water that it is

---

[37] The locus classicus of RA is Dretske (1970), but it's also defended at length by Nozick (1981).

[38] Falvey and Owens make this claim about the possibility that one is thinking that twater is a liquid; and of course they'd be right if one could know that one is thinking that water is a liquid and not that one is thinking that twater is a liquid. Their point, though, holds for this possibility.

dripping. As they characterize it, the theory states the following necessary condition for knowledge:

> If q is a relevant alternative to p, and S's belief p is based on evidence compatible with q's being the case, then S does not know p.

Clearly, one's evidence is compatible with its being the case that one is thinking of some substance distinct from water that it is dripping, for the simple reason that one typically has no evidence at all for one's belief that one is thinking that some water is dripping; so RA should predict that one doesn't know that one is thinking that some water is dripping, contrary to intuition.

But Falvey and Owens explain why RA doesn't work for privileged access:

> In most cases the character of a person's evidence for his belief is a reliable indication of his susceptibility to error.... In [the case of beliefs about one's own mental states, however,] the inability of a subject to eliminate a relevant alternative does not entail that the subject is liable to error.[39]

They point to a perceived asymmetry between ordinary empirical knowledge and privileged access: only in the case of the former is a subject's inability to eliminate a

---

[39] Falvey and Owens (1994), 117.

relevant alternative an indication of the possibility of error. Why is this so? Because of DH. To see why DH works in this way, we need to understand a little more about RA.

According to Falvey and Owens, the appeal of RA depends on the notion of susceptibility to error. When one's beliefs are susceptible to error, they don't count as knowledge. They illustrate this with Alvin Goldman's (1976) well known barn example:

> Suppose that while driving in the country Tom judges that a certain structure nearby is a barn. Tom is competent in the use of the word 'barn', he has excellent eyesight, the object is in plain view, and it is indeed a barn. On the basis of this description of the situation, there seems no reason not to say that Tom knows that the object is a barn. But now suppose that we are given additional information to the effect that unbeknown to Tom, the area in which he is driving is full of papier-mâché facsimiles of barns, which travelers invariably mistake for genuine barns. In light of this additional information, we would not be inclined to say that Tom knows that the object he is looking at is a barn, even if it is a barn.[40]

RA explains our disinclination to say that Tom knows that he is seeing a barn by appeal to the fact that Tom's evidence for his belief that he is seeing a barn is compatible with his seeing a barn facsimile; but why should this be an

---

[40] Falvey and Owens (1994), 114.

indication that Tom is liable to error? The reason, according to Falvey and Owens, is that if Tom were in an environment full of fake barns, then if Tom's belief that he is seeing a barn were false, he probably would still believe that he is seeing a barn. That is, the compatibility of Tom's evidence with the possibility that he is seeing a fake barn (when the possibility that he is seeing a fake barn is a relevant alternative) shows that the following counterfactual is false:

> If Tom's belief that he is seeing a barn were false, then Tom wouldn't believe it.

To use a notion from Nozick (1981), if Tom were in an environment full of fake barns, then his belief would not be sensitive to the possibility that he is seeing a fake barn, where sensitivity is defined as follows:

> S's belief p is sensitive iff S would not believe p if p were false.

So Falvey and Owens conclude that RA implicitly embodies this feature, and call this better articulated theory RA':

> If q is a relevant alternative to p, and S's justification for p is such that, if q were true, S would still believe p, then S does not know p.[41]

We are now in a position to state how we can know that we're thinking that some water is dripping without knowing that we're not thinking of some other substance that it is dripping, and why we need RA' to do so. If DH is true, then whatever determines that our thought that some water is dripping is about water also determines that our thought that we're thinking that some water is dripping is about water. So let's consider the (admittedly) relevant alternative in which we're on Twin Earth instead of Earth. In such an alternative, we're not thinking that some water is dripping; instead, we're thinking that some twater is dripping. What's more, we can't introspectively distinguish between thinking the one thought and thinking the other. But if we <u>were</u> on Twin Earth instead of Earth, we wouldn't <u>believe</u> that we were thinking that some water is dripping; instead, we'd believe that we were thinking that some twater

---

[41] Ibid., 116. For the sake of discussion, the counterfactual <u>if p were true, q would be true</u> should be read in the following way: q holds in every relevantly nearby p-world. We shall ignore both true and impossible antecedents.

is dripping. Therefore, a key conjunct of the antecedent of RA′ is false in this situation, and this undermines the skeptic's argument that we lack introspective knowledge of content.

By contrast, Falvey and Owens point out that ordinary empirical knowledge seems not to enjoy this limited infallibility. When we have some empirical belief, the existence of relevant alternatives we can't rule out does seem to indicate susceptibility to error, as Goldman's barn example shows. As for the case at hand, relevant alternatives can be gerrymandered in such a way that we can see how it would be possible to believe, but not know, that some water is dripping. By contrast, no such gerrymandering seems possible with privileged access--it seems not to be underminable in this way.

At this point we can see why non-empirical Closure must fail if Falvey and Owens are correct: RA′ and non-empirical Closure are mutually incompatible, and Falvey and Owens appeal to RA′ to explain why introspective knowledge of content is compatible with externalism.

So much for their solution. Now we need to see why it doesn't work. I have two objections to it: first, they haven't shown that the sensitivity of privileged access reflects its epistemically relevant features; second, their denial of Closure is suspect. I explain each in order.

Is it epistemically relevant that we wouldn't have water thoughts were we on Twin Earth? Perhaps, but Falvey and Owens offer no reason for thinking so. They claim that there's a key difference between privileged access and most empirical beliefs: with empirical beliefs only does one's inability to rule out a relevant alternative indicate susceptibility to error. But this isn't correct, as the following case shows.

Consider, not thoughts that one is thinking that some water is dripping, but just thoughts that some water is dripping. In the nearest worlds in which one is in a twatery environment, one's thought that some water is dripping is false--but it's also a thought one wouldn't have in that world. So one's belief that some water is dripping is sensitive to that particular relevant alternative. Of course, one's belief isn't sensitive to all relevant

alternatives; but the point is that the sensitivity of a belief to some relevant alternative alone doesn't indicate that the belief isn't susceptible to error. It seems that one's belief that some water is dripping can be as unjustified as you please, and it would still be sensitive to the alternative in which one is on Twin Earth thinking that some twater is dripping. That is, that sensitivity rules out this alternative is arguably epistemically irrelevant--it tells us nothing about whether or not the believer is susceptible to error by virtue of having this belief. Why, then, should we believe that privileged access's sensitivity to the very same alternative (and relevantly similar ones) tells us anything about whether we are susceptible to error by virtue of having beliefs about our own mental states?

Falvey and Owens's denial of Closure is another problem. They are welcome to deny non-empirical Closure--it has by no means been established--but RA' is not only inconsistent with non-empirical Closure, it is also inconsistent with Closure simpliciter. The reason is simple. If RA' is true, one can know what one is thinking

that some water is dripping even if one doesn't know known consequences of one's thoughts--for example, that one is not a brain in a vat in a waterless world. But because they claim that this can't be known at all, much less known non-empirically (see Section 6), they must deny Closure.

Closure, however, iscrucial to the plausible view that we can add to our knowledge by using deduction. Everything we know about reasoning tells us that if someone knows both p and that p implies q, she knows q as well; this is because, upon sufficient reflection, she can deduce q from her other knowledge. The External-World Paradox is hard just because it forces us to choose between denying the plausible view that we have knowledge and denying this plausible principle. One can't just deny it; one must explain why, given that it's false, it seems true. So Falvey and Owens solve one paradox, only to fall into another. They have no explanation for why Closure seems to hold but fails to; so their second solution to the paradox fails because it depends on there being a solution to the External-World Paradox, a solution they don't entertain.

I conclude, then, that denying Closure and relying on the sensitivity of privileged access won't solve the paradox of externalism and privileged access.

## 7. Contextualist solutions

So far as I know, no one has offered a contextualist solution to the paradox in print. But contextualism is an interesting and possibly important theory with obvious *prima facie* implications for the paradox, so it's worth discussing these implications to see whether they provide enough material for a solution. Following arguments of Schiffer (forthcoming and lectures), I will argue that they don't, for two reasons: first, that contextualism is pretty clearly false; second, that even if it were true, it wouldn't solve the paradox. This will become clear in three related objections to contextualism.

Contextualism in epistemology is an attempt to solve the External-World Paradox by appeal to contextual factors involved in knowledge claims. In essence, contextualism amounts to the claim that the truth conditions of knowledge claims, hence the propositions they assert, change from context to context. This is because sentences of the form *I*

know p don't express propositions independent of some contextually specified standard for knowledge: to say that someone knows something is to say that she knows something relative to some or another standard. These standards may shift, so we can't say what proposition a knowledge claim expresses independently of the standards in force in a given context.

Implicit contextual standards are familiar in natural language. For example, when we say "It's snowing," we're typically making implicit reference to some location. (When someone asks about the weather in New York and I say, "It's snowing," I am saying that it's snowing in New York.) It's also arguable that when we say, "Anthony Mason is not tall," we're making implicit reference to a population: Anthony Mason is not tall for a professional basketball player (there are significantly many taller players), even though he is tall for an American (being well over six feet tall). Similar remarks hold for such qualitative predicates as "big", "fast", "thin", and many others.

Contextualists would argue that knowledge is implicitly contextual in just this way. They contend that knowledge

claims make implicit reference to standards, in just the way that claims of flatness and certainty (to pick two salient examples) make implicit reference to standards. Some things are flat enough for a given purpose, so these things count as flat when used for those purposes; outside those contexts, they may not count as flat. One may also be certain enough for some purposes but not others. Knowledge works similarly: to know that p is to have evidence good enough for a given purpose.

Different contextualists offer different explanations of what it is for someone to have good enough evidence. David Lewis, for example, has it that knowing p is having evidence sufficient to rule out every possibility in which not-p, but restricts the range of 'every' to more or less than all of the possibilities on the basis of a variety of conversational rules.[42] Another contextualist, Keith DeRose, doesn't offer an analysis of knowledge claims per se, but posits a conversational rule which requires that the beliefs which one claims to are knowledge be sensitive; in

---

[42] See Lewis (forthcoming).

this way, for some propositions p, one can count as knowing p or not depending on whether one is claiming to know p.[43]

How does this affect the External-World Paradox? Let's recall it:

(A1) I know p.

(A2) If I know p, I know not-q.

(A3) I don't know not-q.

Let p be "I have hands", and let q be "I am not a handless brain in a vat". Let's suppose that (A1) is true. Then if (A3) is true, I don't know that I am not a handless brain in a vat, which would force us to deny (A2). This would be an unhappy position to be in, because we'd be denying an extremely plausible principle. If contextualists are correct, though, (A3) doesn't express a proposition independently of context: in contexts where standards are fairly low, I can count as knowing that I am not a handless brain in a vat. But in contexts the skeptic uses to assert (A3), the standards are as high as can be: in those contexts, I don't count as knowing that I am not a handless brain in a vat. By those standards, of course, I don't know

---

[43] See DeRose (1995).

that I have hands. Therefore, the contextualist can offer a solution to the External-World Paradox which doesn't depend on denying Closure, and which doesn't require her to categorically deny either (A1) or (A2). In ordinary contexts, (A1) is true and (A3) false; in skeptical contexts, (A1) is false and (A3) true. (Presumably, there are no contexts in which all three claims are true.)

How would a contextualist solution to the External-World Paradox apply to the case of privileged access? IMA and its corollaries state that no one can know non-empirically the externalistic consequences of what one knows one thinks. But if the standards for knowledge shift from context to context, it may be that those who appeal to IMA or its corollaries are implicitly relying on distinct contexts to make the separate claims true. The solution would assert that we know what we're thinking non-empirically in ordinary contexts, but not in skeptical ones; and when we know what we're thinking non-empirically, we also know its externalistic consequences non-empirically. When in skeptical contexts, we require empirical investigation to know that externalistic consequences

obtain; but by non-empirical Closure, we also need empirical investigation to know what we're thinking in those contexts.

This is a tricky solution. It doesn't categorically deny any of any of IMA's or its corollaries's premises, but one or another premise of each will be false depending on the context, and no context will allow them all to be true together. Will it work?

Of course, it will only if contextualism is true, and there are compelling reasons for thinking it isn't. There are three objections, each of which seems decisive.[44] First, contextualism requires speakers to make implicit reference to contextual features of knowledge claims, which requires speakers to know what features these are; but speakers seem to have no such knowledge. Second, if contextualism is true then skeptics not only don't know what they mean by their own sentences, but also have systematically false beliefs about what these sentences mean. Third, if contextualism is true, then skeptics don't even know what they're thinking when they think through the

---

[44] Schiffer offers the first two of these objections in his (forthcoming); he offers the third in lectures.

propositions expressed by (A1)-(A3), which conflicts with a view we have assumed since Chapter II, viz., that they do know what they are thinking, and indeed, know it non-empirically.

Let's see how the first objection works. What makes it plausible that claims of tallness have implicit contextual parameters is that speakers can generally make them explicit when called upon to do so. Anthony Mason is tall for an American, though not for a professional basketball player--and anyone who knows what "tall" means can easily explain why the sentence "Anthony Mason is tall" could be false in a sports context but true in the context of a national height survey. These parameters are typically implicit for an obvious reason: it's a pain to constantly have to make them explicit. But for a parameter to be contextually implicit (as opposed to, say, psychologically implicit) just is to say that it's shared knowledge between speaker and audience that the parameter is in place.

This is typically not the case with the allegedly implicit contextual parameters of knowledge claims. If a speaker claims to know that she has hands, it's by no means

clear that she's making any reference at all to shared knowledge that some set of standards is in force. When asked to supply the missing parameter, most if not all speakers would be at a loss for an answer. The question, "By what standard or set of standards are you claiming to know that you have hands?" would likely be met with puzzlement. (Compare: "Anthony Mason is tall by what standard of height?" "It's snowing where?") This is a good reason for thinking whatever implicit contextual parameters there are in knowledge claims, they're not standards for knowledge.

Now, I'm not claiming that we couldn't have a predicate which conformed to the requirements which contextualists allege to hold for 'knows': to use David Lewis's definition, one which is used with implicit contextual parameters, and which applies to an individual i and a proposition p just in case (a) i believes p, and (b) i has evidence e for p such that in context c, e eliminates every possibility in which not-p. It's just that this predicate isn't 'knows' as we use it, and doesn't pick out the knowledge relation as we understand it. One can't solve the External-World Paradox

simply by replacing 'knows' with a homophonic predicate which solves the paradox by definition.

The second objection is also serious. Recall that to solve a paradox is not simply to point out the offending proposition(s); it is also to plausibly explain why the paradox seemed paradoxical in the first place. If contextualism is true, however, the only possible explanation they can offer for why the External-World Paradox seemed paradoxical is unconvincing. If contextualism is true, then (A1)-(A3) should be paradoxical precisely to the extent that the following sentences are paradoxical:

(1) Derek Harper is tall.

(2) If Derek Harper is tall, then Anthony Mason is tall.

(3) Anthony Mason is not tall.

Unfortunately, (1)-(3) are obviously not paradoxical. (1) and (3) are false and true respectively when we compare Derek Harper and Anthony Mason to other professional basketball players, but are true and false respectively when we compare them to the general American population. Anyone who understands "tall" knows this. By parity of reasoning,

then, (A1)-(A3) should not seem at all paradoxical. The only explanation for the External-World Paradox's seeming to be paradoxical, then, is that those to whom it seems paradoxical not only don't know which propositions are expressed in (A1)-(A3), but are also mistaken about which propositions are expressed.

The skeptic should know that in uttering (A1) and (A3), she is expressing distinct relations by her respective uses of 'knows', just as she knows that in uttering (1) and (3), she is expressing distinct properties by her respective uses of 'is tall'. But this is precisely what the contextualist must claim the skeptic doesn't know. What's more, the contextualist must claim the skeptic believes, mistakenly, that 'knows' expresses the same relation in (A1) and (A3), even though she doesn't believe that 'is tall' expresses the same property in (1) and (3).

It should by now be easy to see how the third objection works. If contextualism is true, then the skeptic is not only wrong about what she means by her words, she is also wrong about what thoughts she's thinking. When the skeptic asserts (A2) and (A3) while denying (A1), for example, she

believes the contents of the first two and believes the negation of the content of the third. Unfortunately, the skeptic believes falsely that (A1)-(A3) are mutually inconsistent, which implies that she doesn't know the true content either of (A1) or of (A3). The reason is that she not only can't specify the implicit parameters involved in each proposition, but also mistakenly identifies them. The only way to make (A1)-(A3) seem paradoxical is to conflate distinct contextual parameters; and the only way to conflate distinct contextual parameters is to have false beliefs about one's own thoughts. (Indeed, imagine what it would take to believe that thinking through (1)-(3) is paradoxical: one would have to falsely believe either that one is thinking that Derek Harper is tall for an NBA player or that Anthony Mason is not tall for an American.)

Of course, it's conceivable that skeptics don't know what they're thinking when they assert (A2) and (A3) while denying (A1); but it's not very plausible, given the assumption that we all have privileged access. Skeptics, in general, are extremely intelligent and sophisticated people; is it plausible to think that they regularly get it wrong

about their own thoughts, especially when it comes to thoughts about a subject matter at which they are expert? I think not. Therefore, we can't accept contextualism as a solution to the External-World Paradox. But if we can't accept it as a solution to the External-World Paradox, we can't accept it as a solution to the paradox of externalism and privileged access either.

## 8. Conclusion

We've seen that none of the solutions discussed here are satisfactory. What's worth noting about them, though, is that all these solutions take for granted some notion of non-empiricality: Davidson and Heil accept first-person authority, Burge basic self-knowledge, Falvey and Owens a priori knowledge. Only Falvey and Owens have examined non-empiricality with an eye to whether it supports an argument for (IT), but as we've seen, their solutions are unsatisfactory.

In the next chapter, we'll take a good look at the notion of non-empiricality, to see whether it really does support the best arguments for (IT).

# VI

# SOLVING THE PARADOX

## 1. Introduction

In Chapter V, we saw that none of the previously offered solutions to our paradox were satisfactory. For the most part, this was because the solutions offered didn't address the best arguments for the paradox; but even the one which addressed these arguments offered an implausible response to them.

In this chapter, I use a slightly different approach to the paradox. Rather than simply picking a premise and arguing against it, I examine IMA and its corollaries afresh. Throughout Chapters IV and V, I accepted uncritically the notion of non-empiricality employed by the best arguments for (IT). Now is the time for us to critically examine this notion, and we'll see that upon examination, we get not one but three distinct notions. Our choice of notion, in conjunction with some version of externalism, yields different possible solutions to the

paradox, and as it happens there is no one notion according to which all the premises of IMA and its corollaries are true.

Of course, showing that the best arguments for (IT) are unsound is only half a solution to the paradox of externalism and privileged access; the other half is explaining why (IT) seemed plausible to begin with. It can't be because of IMA, which is unsound, so we must find some other explanation. I'll suggest two distinct but related explanations for why someone might have accepted (IT): first, uncritical acceptance of a single notion of non-empirical knowledge; second, uncritical acceptance of the view that privileged access consists in some kind of privileged knowledge of the propositions one is thinking.

I'll have more to say about these two explanations later on. First, though, let's take a fresh look at IMA and its corollaries.

## 2. IMA and its corollaries restated

For clarity's sake, let's restate IMA and its corollaries.

**The Improved McKinseyan Argument (IMA)**

**(F1)** If we have privileged access, then Nathan Salmon knows non-empirically that he is thinking that water is wet.

**(F2)** If externalism is correct, then Nathan Salmon knows a priori, hence non-empirically, that his thinking that water is wet entails externalistic fact E (where E is an appropriate externalistic content-determiner).

**(F3)** If x knows non-empirically both p and that if p then q, x can know q non-empirically.

**(F4)** Therefore, if we have privileged access and externalism is correct, then Nathan Salmon can know E non-empirically.

**(F5)** But no one can know E non-empirically.

**(IT)** Therefore, privileged access and externalism are incompatible.

**The Improved Skeptical Argument (ISA)**

Given IMA, an appropriate externalistic content-determiner is the proposition that Nathan Salmon is not thinking of some indistinguishable substance distinct from water that it is wet. Nathan Salmon arguably can't know this proposition non-empirically, but he does know non-empirically that his thinking that water is wet entails it; so privileged access and externalism are incompatible.

**The Anti-Putnamian Argument**

Given IMA, the appropriate externalistic content-determiner is the proposition that Nathan Salmon is not

> a (Putnamian) brain in a vat. Nathan Salmon arguably can't know at all, so a fortiori doesn't know non-empirically, that he's not a brain in a vat, but he does know non-empirically that if he's thinking that water is wet, then he's not a brain in a vat; so privileged access and externalism are incompatible.

## 3. Non-empiricality refined

Let's begin with a crude notion of non-empirical knowledge: to say that something is known non-empirically just is to say that it is known in a way that's independent of empirical investigation. This notion is far from satisfactory--independence in what sense? what exactly is empirical investigation?--but it's a good starting point.

Non-empiricality, so understood, is indistinguishable from an informal notion of a prioricity, one used by some discussed in this essay (e.g., McKinsey 1991 and Falvey and Owens 1994). Later on, we'll see how more rigorous notions of a prioricity can play a role in a better understanding of non-empiricality.

When characterizing privileged access in Chapter II, I wrote of the apparent "non-empirical authority" we have about what mental states we're in. I cashed this out in terms of self-warranted belief: beliefs about our own mental

states which are justified simply by virtue of our having those beliefs. Obviously, if a belief is self-warranted, it's not warranted by virtue of being based on empirical evidence; so our authority over our own mental states doesn't seem to depend on empirical investigation to that extent.

I also claimed, in Chapter IV, that our knowledge of externalism seems not to depend on empirical investigation, because it is arrived at through a priori philosophical analysis of the notion of propositional content. This suggests that externalism, if true, is knowable non-empirically, by virtue of being knowable a priori. What's more, the externalistic dependence claims put forward by the three externalisms described in Chapter III are also arrived at through philosophical analysis, so are also plausibly knowable non-empirically if knowable at all.

But for a paradox whose force compels some to question externalism, others to question privileged access, and still others to question Closure, its motivation ought not rest on such a vague, amorphous notion as this initial one of non-

empiricality. It is our task here to examine the notion more closely, to see what fruits it may yield.

### 3a. Non-empiricality and a prioricity

We saw that our initial characterization of non-empiricality was indistinguishable from an informal notion of a prioricity. But surely we can do better than this. What does a prioricity amount to?

There are far too many notions of a prioricity for it to be useful to catalogue them. But Hartry Field has put forward two notions of a prioricity which capture what most of us have in mind when using the term 'a priori':

> A principle is <u>weakly a priori</u> if it can be known or justifiably believed on a basis other than empirical evidence for it; <u>strongly a priori</u> if in addition it is empirically indefeasible, that is, if there is no possibility of undercutting our apparent knowledge of it or justified belief in it by empirical evidence.[1]

These two notions will be quite useful, but we must first note a clarification Field makes of the notion of defeasibility. Field notes that there are ways of undermining knowledge or justification which do not seem to

---

[1] Field (forthcoming), 1.

seriously call into question the a priori status of intuitively a priori claims:

> No a priorist would deny that empirical evidence that one has carefully checked a proof of p and found it faultless, or that clever logicians had done so, counts as some sort of empirical evidence for a logical principle p. Correspondingly, I can claim to know a priori that p while admitting that my confidence should be undermined by evidence that I was on drugs when I advanced the proof that p and checked it, or evidence that logicians unanimously agree that my argument that p is faulty.[2]

Thus, Field distinguishes between two kinds of empirical defeat: one kind is evidence which "outweighs the non-empirical basis we had for making the claim," which he calls primary undermining evidence; the other kind is evidence which shows "that we did not after all have the non-empirical basis we thought we had for making the claim," which he calls secondary undermining evidence.[3] The only kind of evidence relevant to claims of strong a prioricity, then, is primary undermining evidence.

If we identify non-empirical knowledge with weak a priori knowledge, we capture much of what we would hope to capture about the non-empiricality of our knowledge of our

---

[2] Ibid., 4.

[3] Ibid., 5.

own mental states and of externalism. If by 'on a basis other than empirical evidence' we mean 'not on the basis of empirical evidence', then the propositions we believe in having privileged access are definitely weakly a priori: they are known or justifiably believed on no evident basis at all. And belief in externalism is justified on a basis other than empirical evidence for it, if, as we assumed in Chapter III, it is justified by a priori philosophical analysis.

But weak a prioricity may not capture all of what we wish to capture using the notion of non-empiricality. As I noted in Chapter II, we typically regard empirical evidence to be irrelevant to the question of whether the beliefs we have about our own mental states (in having privileged access) are true, so it's hard to see how empirical evidence could undermine our claims to know we're in those states. And we don't seem to regard particular empirical facts to be relevant to the question of whether externalism is true either--it's not a thesis whose truth seems contingent on some particular empirical facts holding. We can't capture these features with the notion of weak a prioricity. We

need strong a prioricity--weak a prioricity fortified by indefeasibility by primary undermining evidence--to capture them.

This gives us two good ways of cashing out the notion of non-empiricality used in IMA and its corollaries: weak and strong a prioricity. Strong a prioricity captures features which seem to hold for both privileged access and externalism, but weak a prioricity does qualify as "non-empirical" and so deserves consideration as well.

### 3b. Externalism and defeasibility

I pointed out that privileged access seems to be strongly a priori. Externalism, however, complicates this intuition. For example, I might come to believe that a certain state I'm in lacks content, on the ground that I fail to bear some appropriate externalistic relation to anything. So I might, say, come to believe that the state I'm in which I would express by using the sentence "I think I'm talking to Hillary Clinton" lacks content, because I believe that I bear no appropriate externalistic relation to

the referent of "Hillary Clinton", perhaps because I suspect I'm hallucinating (and that the name is vacuous).[4]

Clearly, there might be evidence that "Hillary Clinton" is vacuous, and that everyone who ever had experiences as of "her" was hallucinating. I might also become aware of this evidence, even as I believe that I think I'm talking to Hillary Clinton. Does this evidence undermine my belief? If so, then externalism shows that privileged access is not strongly a priori, contrary to intuition.

The answer to this question calls for some delicacy. Normally, we count evidence as undermining a claim if it suggests the claim is false--that is, if it counts as a reason for believing it false. But the evidence appealed to doesn't count as a reason for believing that my belief that I think I'm talking to Hillary Clinton is <u>false</u>. That's because if Hillary Clinton doesn't exist, I bear no externalistic relation to her, and so not only do I not think I'm talking to Hillary Clinton, I also don't <u>believe</u> that I think I'm talking to Hillary Clinton. If my lower-

---

[4] I owe this example to Stephen Schiffer. Keep in mind that in this case I'm assuming something like strong externalism; I'll discuss other externalisms in good time.

order state lacks content by virtue of my failing to bear the appropriate externalistic relation, my higher-order belief likewise lacks content, for the same reason. Indeed, this is the lesson of the Davidson-Heil Thesis, discussed in Chapter V.

This kind of undermining, then, can't be normal, so can't be a normal kind of empirical defeat. Nevertheless, it seems clearly to be some kind of empirical defeat. Thus, externalism forces us to distinguish between two kinds of primary undermining evidence. The first kind we might as well call falsifying, because it counts as a reason for thinking the claim in question is false. The second kind I'll call emptying evidence, because it counts as a reason for thinking a state (or sentence) lacks content altogether.

### 3c. IMA re-examined

What are we to say of IMA, given these distinctions? Let's take each kind of non-empiricality in turn, examining each key premise of IMA.

One explanatory remark is called for before this examination, however. Throughout I will accept (F3), non-empirical Closure. My reasons for doing so are two. First,

I see no good reasons for denying non-empirical Closure on any account of non-empiricality; second, I am uncomfortable with denying any closure principle in the absence of non-question-begging counterexamples. One trouble with denials of epistemic closure in the literature is that they seem to be motivated by the desire to avoid skeptical conclusions, but these denials are appealed to in order to explain why these conclusions don't follow, which seems to me paradigmatically question-begging. I could do the same here: I could simply deny (F3) and appeal to "intuition" that some antecedent instance of (F3) is true while its consequent instance is false. In this context, though, such a move would beg the question.

(i) Weak a prioricity

It's clear that privileged access is weak a priori knowledge, and it's also clear that externalism can be known or at least justifiably believed on a basis other than empirical evidence. Thus, externalistic conditionals which follow from each version of externalism are also so knowable or justifiably believable. If non-empirical knowledge is

weak a priori knowledge, then, (F1) and (F2) are each true. Given (F3), then, (F4) must also be true.

What of (F5), though? One might think that (F5), too, is true: no one can have a claim to know some externalistic consequence on any basis other than empirical evidence. But why must this be so? Given that the inference from externalism and his knowledge of his own thoughts to E is available to him, why shouldn't Nathan Salmon have adequate non-empirical justification in believing E? Indeed, why shouldn't we accept that Nathan Salmon knows E on a non-empirical basis?

One objection might run: because if he could know E on a non-empirical basis, skepticism would be false a priori, and while skepticism may well be false, it surely isn't false a priori. But this argument is far from conclusive. At best, that Nathan Salmon knows E non-empirically conflicts with the claim that he can't know E at all, much less non-empirically--and this conflict is nothing more than an instance of the External-World Paradox. Perhaps there is a solution to this paradox, but I wouldn't have thought that explaining why externalism is compatible with privileged

access should require solving some other paradox. (Besides, I don't know why skepticism, if false, shouldn't be false a priori. Should it instead be false only as a matter of fact? Why?)

But this objection hints at a deeper one. Skepticism is based on the intuition that any evidence we might have for a claim can be undermined by skeptical considerations. So the appeal to skepticism here is not irrelevant: it seems that while privileged access and our a priori knowledge of externalism are both insusceptible to skepticism, our knowledge of externalistic consequences surely is so susceptible--this knowledge is paradigmatically empirical. Therefore, (F5) must be true, because skeptical considerations plainly can undermine our claims to know externalistic consequences. If (F1)-(F3) are all true, however, we still have a paradox.

This deeper objection is more interesting, but it appeals to a notion of non-empiricality stronger than weak a prioricity--that is, strong a prioricity. I'll have more to say about this objection when I discuss strong a prioricity given emptying evidence.

(ii) Strong a prioricity given falsifying evidence only

As noted earlier, privileged access seems to be strongly a priori knowledge if we include only falsifying evidence. So intuition supports the view that (F1) is true given this notion.

The case of (F2) is more complicated. Strong externalism, according to which thinking a thought about water depends on having causally interacted with water, seems clearly defeasible by falsifying evidence if we allow that someone could have a thought about water even if there never has been any water--that is, if we accept Burge's (1982) weaker version of externalism. Thus, (F2) is false on this notion, given strong externalism.

What of weak externalism? I confess to find it impossible to imagine how someone could have thoughts about water even though there have been neither any water in her environment, nor have any waterhood hypotheses in her community. If neither of these things is the case, what could make her 'water' states and utterances determinately about water and not some indistinguishable non-watery substance (or nothing at all)? If she discovered evidence

that there never had been any water and that there had been no hypotheses of waterhood in her community, I'm inclined to think that she shouldn't doubt that if externalism is true, then her thinking water thoughts entails what her evidence suggests isn't so; rather, I think she should doubt that she has water thoughts at all. Accordingly, I think that weak externalism is strongly a priori, and that the externalistic conditionals it issues are strongly a priori if we include falsifying evidence only.

But this is controversial. I'm suggesting here that, ultimately, it's not coherent to doubt weak externalistic conditionals. I'm sure that some would bristle at this: individualists, in particular, would claim that it is perfectly coherent to believe that one is thinking water thoughts while denying both that there is any water and that there have been waterhood hypotheses in one's community.

Stephen Schiffer (in conversation) offered an illustrative analogy. By using a priori philosophical analysis, a functionalist about mentality might identify pains with neural states with such-and-such functional roles; but if we discovered that there weren't any neural

states with those roles, we wouldn't come to believe that there weren't any pains, but would instead deny the identification of pains with such states. Likewise: if we discovered not only that there wasn't any water in our environment, but also that there haven't been any waterhood hypotheses in our community, we might not come to believe that we don't have any water beliefs; we might instead attack the inference from having water beliefs to their weak externalistic consequence.

I don't know how to resolve this issue, and I won't try, now that I've stated my own view of it. But we should keep in mind that which side we take will give us a different response to IMA. If we accept that we can know externalistic conditionals non-empirically on this notion, the (F2) is true. But accepting (F2) requires us to claim it isn't coherent to believe that one is thinking that water is wet while denying its weak externalistic consequence. And if we must do this to accept (F2), it follows that (F5) is false on the current notion of non-empiricality: there is no evidence which could give us a reason for thinking that someone believes a weak externalistic consequence falsely.

This brings up the deeper objection raised in the subsection on weak a prioricity. It surely seems that our knowledge of externalistic consequences can be undermined by falsifying evidence--indeed, if they couldn't be, radical skepticism would be an incoherent position, and it seems perfectly coherent (if controversial). But there are two ways of responding to this objection here. First, the force of this objection might depend on whether it's coherent to doubt weak externalistic conditionals; if so, then the objector, while accepting (F5), would have no reason to accept (F2). Second, to say that knowledge of externalistic consequences can't be undermined by falsifying evidence isn't to say that it can't be undermined at all. It might be possible to undermine it by emptying evidence. A radical skeptic needn't be so picky about what kinds of empirical defeat undermine knowledge, if she can show that some kind of undermining is possible. As we'll see, our knowledge of weak externalistic consequences can be undermined by emptying evidence, so we can explain the intuition that radical skepticism is a coherent position by pointing out

that externalism plus privileged access doesn't rule out the possibility of this kind of empirical defeat.

All this is predicated on the assumption that it's not coherent to deny weak externalistic conditionals. If we accept that it's coherent to believe that one is thinking that water is wet while denying its weak externalistic consequence, then we have no reason to doubt (F5). But we would then be forced to deny (F2), because weak externalism would be true, even though no one knows weak externalistic conditionals non-empirically.

We can't have it both ways: if non-empiricality is strong a prioricity given falsifying evidence only, then either (F2) is false or (F5) is false; they can't both be true on this notion of non-empiricality.

Super-weak externalism is easy to handle given weak externalism. Because it follows from weak externalism, whatever holds for the latter holds also for the former.

(iii) Strong a prioricity given emptying evidence

This kind of non-empiricality yields interesting results about IMA. First of all, we've noted that beliefs about our own mental states, even in having privileged

access, might be defeasible by primary undermining evidence if we allow that such evidence be emptying: I might come to believe that the state I'm in such that I would express it using the sentence 'I think I'm talking to Hillary Clinton' lacks content, and on the basis of evidence that Hillary Clinton doesn't exist. Thus, given strong externalism, (F1) is clearly false on this notion of non-empiricality.

Given weak externalism, (F1) is also clearly false given emptying evidence. There might be evidence that there has been neither any water in one's environment nor any waterhood hypotheses in one's community; this evidence could undermine one's confidence that one was in fact thinking water thoughts, without actually falsifying one's belief that one was thinking those thoughts. It would instead give one reason to believe one's 'water' states lack content.

What we say about (F2) is similar to what was said in (ii) above. If it's coherent to believe one is thinking water thoughts while denying its externalistic consequence, then (F2) will be false given emptying evidence--emptying evidence suggests that the consequents of externalistic conditionals are false. Otherwise, (F2) is also true.

What, then, of (F5)? Again, if it's coherent to believe that one can have water thoughts while denying their externalistic consequence, then (F5) is true. If it isn't, then (F5) is false.

So (F1) is clearly false given emptying evidence, even on weak externalism, while (F2)'s and (F5)'s truth depends on the coherence of denying weak externalistic conditionals.

As for super-weak externalism: the notion of emptying evidence doesn't apply. As noted in Chapter III, there's no way for a state to lack externalistic content given super-weak externalism. So (F1) and (F2) both seem true if given this notion of non-empiricality and super-weak externalism. Furthermore, (F5) seems plainly false: there's no evidence which would show that a belief state which expresses a super-weak externalistic consequence lacks content.[5]

---

[5] A clarification is necessary here. One might find evidence which, due to some mistake, leads one to believe that a certain state lacks content. For example, one might believe strong externalism, and believe that by virtue of there not being any water that one's state lacks content. But this ought not count as a possible kind of empirical defeat. For a proposition to be a possible defeater of one's knowledge of an externalistic consequence, it must be such that <u>if it were true</u>, then either one would believe that consequence falsely, or one's belief state would express no proposition at all.

Again, the deeper objection arises, this time with more force: surely evidence can undermine one's knowledge, say, that there isn't a non-watery substance indistinguishable from water instead of water in one's environment?

The response to this objection is in two stages. First, I have the intuition that if it turned out that some substance like water were discovered to exist in our environment instead of (what is actually) water, we would be inclined to think of it as water. Indeed, if we didn't have access to individuating chemical analyses, we wouldn't have the resources to actually discover that this substance wasn't identical to what we originally thought of as water. And, as I noted in Chapter III, given our chemical analyses, if we discovered evidence that these analyses were mistaken so that twater instead of water were in our environment, we'd be inclined to say not that there wasn't any water but rather that water isn't $H_2O$. So such evidence wouldn't be evidence we regard as defeating.[6]

---

[6] Of course, if such evidence turns out to be correct, so that indeed there was twater instead of water in our environment, we wouldn't have the relevant belief to be defeated; so again, the evidence wouldn't be defeating.

But there is a deeper problem. We're not inclined to think that if we were (Putnamian) brains in a vat, then there would be water in our environment. Unfortunately, on externalistic grounds we might be forced to admit that our 'water' thoughts might indeed be about, say, states of the supercomputer controlling us. But evidence that we are brains in a vat is not evidence we would regard as showing that water is some kind of computer state: we would regard even super-weak externalistic consequences as false in this circumstance.

This brings us to the second stage. If this situation were to arise, where we're presented with evidence that we're brains in a vat, we would have evidence that our 'water' thoughts are not about water, but instead about states of a supercomputer. Thus, (F1) would be false, even though our higher-order beliefs wouldn't be falsified. Rather, we would have reason to think that both our lower-order thoughts and higher-order beliefs had what we might call pathological content. Evidence that our states have pathological content is like emptying evidence, although it's not itself emptying evidence.

The case of brains in vats, then, shows us a new but related kind of empirical defeat, one that undermines our faith in (F1).[7] It shows the exception to the plausible view that no evidence would undermine our faith in super-weak externalistic consequences.

### 3d. Summary of results

We've seen that on no interpretation of 'non-empirical' is IMA sound: given weak a prioricity, (F5) is false; given strong a prioricity and falsifying evidence, either (F2) or (F5) is false; given strong a prioricity and emptying evidence, (F1) is false on strong and weak externalism, while (F5) is false on super-weak externalism except in the case of the Anti-Putnamian argument, in which case (F1) is false because of possible evidence of pathological content.

Because there is more than one possible sense of 'non-empirical', it's tempting to diagnose the problem with IMA

---

[7] It's worth noting that evidence that our states have pathological content might hold for both strong and weak externalism as well. So (F1) would be false given such evidence on both weak and strong externalism, while (F2) would still be false on strong externalism and (arguably) true on weak externalism.

as equivocation on 'non-empirical'.[8] But I don't think there is any equivocation here. The issues are too complex to involve anything so straightforward as a logical fallacy. Instead, the main problem has to do with this complexity: it's so hard to think through these issues without getting entangled in them.

Nevertheless, I'll have a few things to say about what makes (IT) prima facie compelling in the next section, things which should make a little clearer the motivation for believing that externalism and privileged access are incompatible.

## 4. Diagnosing belief in (IT)

One major reason for believing (IT) has to do with having an unclear sense of non-empiricality. Numerous writers on the problem of externalism and self-knowledge appeal to an unexplicated notion of a prioricity, and have used it in order to justify IMA-style arguments. As we've seen, though, critical scrutiny of this notion doesn't

---

[8] Diana Raffman (1994) urged a line close to this.

support IMA. So uncritical acceptance of an unexplicated notion of a prioricity can surely lead one to accept (IT).

But there's another reason for this acceptance. Many think of privileged access as privileged knowledge of the contents of one's thoughts. This thought is harmless in the context of individualism, because individualistic content is always discriminable: on individualism, no two thoughts can differ in content but have the same inferential role. Once we accept even a very mild externalism, however, the thought becomes problematic. One reason has to do with Schiffer's toothache case: I don't know non-empirically that I have a toothache, because evidence might arise which suggests that I don't have teeth. That is, I don't have privileged access to the proposition that I have a toothache.

Of course, in another sense I surely do have privileged knowledge of my own toothache: I have privileged access to my own sensory qualities, and so can know non-empirically that I am in a pain state which I am describing as a toothache. But here we're distinguishing between having privileged knowledge of the state I'm in and having privileged knowledge of its propositional content.

The same seems to hold for intentional states. I may lack privileged access to the proposition that I am thinking that water is a liquid, depending on what non-empirical knowledge amounts to. But it's strongly intuitive that I have privileged access to the state I'm in which I describe as a state of thinking that water is a liquid. I know non-empirically that I'm in that state, even if I don't know non-empirically that it expresses the proposition that I'm thinking that water is a liquid.

One might at this point be tempted to think that what this means is that we have privileged access to the narrow contents of our intentional states.[9] But this doesn't mean that at all. For one thing, nothing in this view requires there even to be narrow content; for another, what we would know non-empirically is that we're in a state which we would describe in such-and-such ways, and this knowledge is not obviously knowledge that we're in a state with such-and-such a narrow content. It depends on what narrow content amounts to. (Of course, one could stipulate a notion of narrow

---

[9] Laurence Bonjour (1991) seems to hold a view rather like this one.

content such that this knowledge just was knowledge of narrow content, or that this knowledge was necessary and sufficient for knowledge of narrow content. But I fail to see what would be accomplished by doing this.) In any case, I wouldn't want a view of privileged access to be committed to the notion of narrow content; and this view of privileged access plainly isn't.

This view of privileged access accords well with the intuition that we have non-empirical authority about what mental states we're in, and sidesteps externalism altogether. We can explain the epistemic features of privileged access without falling into paradox. Once we accept this, motivation for (IT) wanes, even though motivations for privileged access and externalism remain as strong as ever.

It's easy to see why one might be tempted to think that privileged access consists partly in knowledge of the propositions one is thinking. As I noted above, individualism may be a hidden assumption; but there is another, less contentious reason. To know what one is thinking is, at least at first blush, to know <u>that</u> one is

thinking such-and-such. We have no reason to deny this. But it's easy to slip from this to the claim that to have privileged access is to have this knowledge in a privileged way. We've seen no good reason yet to accept that view, and some compelling reasons not to; if we can explain how we can have non-empirical authority about what mental states we're in without accepting it, why should we accept it?

Thus, uncritical acceptance that privileged access consists partly in this kind of privileged knowledge is another important reason for acceptance of (IT).

## 5. Conclusion

We've seen that IMA and its corollaries are all unsound, because no notion of non-empiricality supports them. We've also seen that two tempting mistakes can lead one to accept (IT) and put forward an argument like IMA: uncritical acceptance of an unexplicated notion of non-empiricality, and uncritical acceptance of a controversial view of privileged access. Thus, there is now no good reason to accept (IT). Given that we now also have reason for thinking that privileged access is a kind of knowledge which is independent of externalism, we can safely reject

**(IT), the third proposition leading to the (now, merely apparent) paradox of externalism and privileged access.**

# VII

# EXPLAINING PRIVILEGED ACCESS, AND CONCLUSION

In this chapter, I discuss what explains privileged access, and end with a few concluding thoughts. I won't have a lot to say about privileged access, but I hope what I do say will shed a small amount of light on how best to approach the topic.

## 1. Explaining privileged access

### 1a. Naturalism vs. transcendentalism

We might divide accounts of privileged access into two: there are the naturalistic accounts, best represented by David Armstrong (1969), David Rosenthal (1986), and William Lycan (1987); and there are transcendental accounts, best represented by Donald Davidson (1984) and Sydney Shoemaker (1990). (There is also the eliminativist account, best represented by Ryle (1949)--the view that privileged access is really a disguised form of empirical knowledge. On this view, the myth of privileged access arises from the fact

that we typically have much more information about ourselves than we do about others, so we're more reliable and able to use it more easily in making self-ascriptions of mental states than we are about other matters. I'll take it that this view is untenable, because it requires that privileged access be inferential knowledge, which it plainly is not.)

Naturalistic accounts seek to explain the psychological and epistemic features of privileged access (sketched in Chapter II) in naturalistic terms. So for example, we might explain the non-inferentiality of privileged access in terms of an involuntary belief-monitoring mechanism, and explain its non-empirical authority in terms of the reliability of this mechanism.

Transcendental accounts, by contrast, seek to explain the epistemic features of privileged access by appeal to (alleged) conceptual factors, such as the incoherence of the notion that we aren't generally reliable self-ascribers of mental states (Shoemaker), or the dependence of interpretation on this general reliability (Davidson). The psychological features are left unaddressed, but then the main philosophical interest in privileged access is

epistemological, not psychological. The transcendentalist might agree with the naturalist about the existence of some sort of involuntary mechanism, or might assert, with many others, that there is no real difference between being in a mental state and being conscious of it in the way required for privileged access--that is, that consciousness is intrinsic to mentality.

Which approach is best? I'll suggest we take a bit from each.

### 1b. Psychological features: against intrinsicality

It's tempting to think that there is no real difference between being in a mental state and being conscious of it (in the way required for privileged access--henceforth I'll drop this qualification). I noted in Chapter II that if we deny this difference, there's an easy way to explain privileged access's non-empirical authority. But why might one deny it? Rosenthal offers a pithy explanation:

> Consciousness is so basic to the way we think about the mind that it can be tempting to suppose that no mental states exist that are not conscious states. Indeed, it

> may even seem mysterious what sort of thing a mental state might be if it is not a conscious state.[1]

The basicness of consciousness has to do with its felt immediacy, and this immediacy seems best accounted for by supposing that there is no difference between being in a mental state and being conscious of it. The explanation seems to be: if there were such a difference, it would be possible to imagine being in a mental state without being conscious of it, but it's clear that, from the first-person perspective at least, this is not possible.

But, as Rosenthal has pointed out, another explanation is available. Another way of explaining my inability to imagine being in a mental state without being conscious of is by pointing out that imagining being in a mental state from the first-person perspective just is imagining being in a conscious mental state. So even if there is a difference between being in a mental state and being conscious of it, we should expect to be unable to imagine being in a mental state without being conscious of it. Without further

---

[1] Rosenthal (1986), 329.

evidence, there's no way to adjudicate between these two explanations.

But there are reasons for thinking that consciousness is not an intrinsic property of conscious mental states. These reasons come from both psychological theory and from common sense. Cognitive psychology and psychoanalysis make constant appeal to mental states which aren't conscious, and to many which, it seems, can't be conscious. Explanations of such phenomena as visual illusions, semantic priming, and denial all involve essential reference to non-conscious mental states. Methodological difficulties have been the bane of many explanations involving non-conscious mentality, but there's been significant progress in developing techniques for verifying the existence (and structure) of non-conscious mental processes;[2] what's more, much of this research has been remarkably successful. So there is very strong empirical support for the existence of non-conscious mental states, as well as for the idea that consciousness is not intrinsic even to mental states which are conscious.

---

[2] For a discussion of some of the methodological developments, see Ericsson and Simon (1984).

Common sense also supports this view. Rosenthal points out that when

> a headache lasts several hours, one is seldom aware of it for that entire time. Distractions occur, and one pays attention to other things, or just forgets for a bit. But we do not conclude that each headache literally ceases to exist when it temporarily stops being part of our stream of consciousness, and that such a person has only a sequence of discontinuous, brief headaches.[3]

Suchlike cases are legion. So we have both theoretical and common-sense considerations in support of what we might call the extrinsicality of consciousness--its consisting in some sort of relation between the mental state of which one is conscious and some other mental state, perhaps a belief.

Several proposals have been put forward for an explanation of this relation. Armstrong and Lycan each posit a mechanism of "inner perception" which is modeled on ordinary perceptual awareness. Rosenthal posits a mechanism for forming higher-order occurrent intentional states, which he calls higher-order thoughts, by virtue of which we're conscious of our conscious mental states.

---

[3] Rosenthal (1986), 349.

I can't go into detail about the proposals here, but I prefer Rosenthal's proposal to Armstrong's and Lycan's. The trouble with the "inner perception" view is that it's based on an analogy with a process which is itself conscious; so if it doesn't in the end amount to a higher-order-thought proposal, it's hard to see how appeal to a conscious process could be in and of itself be informative.

The higher-order-thought theory is worth briefly spelling out. It says that a conscious mental state is conscious by virtue of its being accompanied by a (roughly contemporaneous and non-inferentially acquired) higher-order thought to the effect that one is in that state.[4] (Of course, the higher-order thought itself needn't be conscious--indeed, normally won't be, and can't be on pain of a vicious regress.)

---

[4] When I say "non-inferentially acquired", I am ruling out non-conscious processes which are described in inferential terms, as in the computation of syntactic structure in sentence processing. I have in mind conscious inferences alone. Rosenthal says in his (1986) that the higher-order-thoughts by virtue of which we're conscious of our conscious mental states are caused (non-inferentially) by them; in later work, he relaxes the causal requirement.

The higher-order-thought theory accounts well for much about conscious mentality that might otherwise seem mysterious. For example, we're often conscious of our sensations more or less coarsely, depending on (among other things) our attention: we taste the nuances of a fine piece of chocolate better when we're paying greater attention to the taste than when we're not. According to the intrinsicality model, we'd have to say, against intuition, that there are different kinds of sensations, some of which involve more nuances than others. The higher-order-thought theory says that the degree of detected nuance is a function of the degree of coarseness of the higher-order thought about the sensation of taste.[5] We can also explain how we can be conscious of a mental state at some times but not others by appeal to this theory, as well as how we can be deliberately as opposed to normally attentive to our own mental states. (The former involves a conscious higher-

---

[5] Rosenthal (1991) suggests that the theory can also explain why education can often affect the way in which one can be conscious of one's own sensations. See his wine-tasting example.

order thought.) These features of mentality are hard to explain on the intrinsicality model.

Rosenthal's theory is not unproblematic. One of the most common objections is that it can't account for the consciousness of creatures which can't think very well.[6] (Interestingly, he anticipated these objections in his (1986); although few seem impressed with his response, little of substance has been said to counter it.) But the research which shows that certain animals have rudimentary capacities for thought is of course based on a comparison with human thought, which is, by contrast, highly sophisticated. Animals may have rudimentary cognitive capacities, but it's question-begging to argue that they thereby lack the requirements for forming higher-order thoughts. What's needed is a principled argument that such-and-such a degree of cognitive sophistication is necessary for thought, and I don't know of any such argument.

Other objections have been offered, but we needn't go into them here. My main point is merely that we should

---

[6] Most recently, Michael Tye (1994) and Fred Dretske (1995) have offered this objection.

prefer extrinsicality to intrinsicality as an account of the non-inferentiality of privileged access.

### 1c. Explaining privileged access's epistemic features

The higher-order-thought theory, however, doesn't seem to have the resources to explain the non-empirical authority of privileged access. Indeed, it would seem to make it mysterious. One might argue as follows:

> If the higher-order-thought theory is correct, then presumably some mechanism underlies my awareness of my own mental states. So, for example, my becoming conscious of my own pains depends on the proper functioning of this mechanism. But if that were so, I could imagine evidence that my higher-order-thought-producing mechanisms are out of whack, hence that although I may firmly <u>believe</u> that I am in pain--indeed, intense pain--I am not really in pain at all. But this doesn't make sense--how could it seem so vividly to me that I'm in intense pain when I'm not?[7]

Thus, the transcendental account of non-empirical authority would seem far better than an account based solely on naturalistic considerations. Davidson, for example, writes:

> There is a presumption--an unavoidable presumption built into the nature of interpretation--that the speaker usually knows what he means. So there is a presumption that if he knows that he holds a sentence true, he knows what he believes.[8]

---

[7] This worry is due to Stephen Schiffer.

[8] Davidson (1984), 111.

Thus, we're forced to presume that speakers generally know what they believe, on pain of their being uninterpretable.

Shoemaker offers a slightly different account:

> I think that it is in fact true that for a large number of mental states we cannot imagine discovering counterexamples to the special authority claim [that is, to the view that we have non-empirical authority about them], even though we can (with some difficulty) imagine discovering counterexamples to the infallibility thesis. This is explicable on the supposition that the mental concepts involved are defined ... in such a way as to make the truth of the special authority thesis constitutive of the mental states in question.[9]

If Shoemaker is right, then the non-empirical authority of privileged access is explained by the fact that the concepts of mentality involved are constituted in such a way that nothing counts as instantiating them unless we're in general authoritative about whether or not we ourselves instantiate them. Thus, it's part of the concept pain, for example, that we're generally authoritative about whether or not we're in pain.

If either of these accounts are correct, then to look to mechanisms for an account for the non-empirical authority of privileged access is simply to look in the wrong place.

---

[9] Shoemaker (1990), 205-6.

The authority we have about our own mental states is derived from features of our practices of self-ascribing them--we simply don't count any state as a pain of ours unless we're in general authoritative about whether or not we're in that state. This may be because we're constrained to in order to be interpretable (Davidson), or because it's constitutive of the concept pain (Shoemaker). But on either account, mechanisms underlying pains and thoughts about them have nothing to do with the explanation of this authority.

### 1d. A mixed account of non-empirical authority

I find the transcendental account hard to believe as the sole explanation of non-empirical authority. What Davidson and Shoemaker seem to be explaining is not so much our non-empirical authority in having privileged access as the necessity of presuming we have it. It's clearly a real feature of us that when we believe we're in some mental state, we're typically right, and part of the explanation of this has to be naturalistic. Something is responsible for our being in those states in the first place, and it's not features of our practices that does it. How could those facts by virtue of which we are in the right states when we

think we are be irrelevant to an explanation of why we're in the right states when we think we are?

But Schiffer's objection is not to be dismissed. It really doesn't seem to make sense that we might seem to ourselves to be in intense pain even though we're not in pain at all, and a naturalistic account of non-empirical authority seems to allow for this possibility.

Part of the trouble is that we do have entrenched self-ascription practices which have nothing to do with our knowledge (as such) of the mechanisms underlying consciousness. These practices reflect the amount of interest we have in our mental states in themselves, apart from our being conscious of them. As Rosenthal puts it,

> when mental states are not conscious, our interest in knowing about them is greatest with propositional states, less with emotions, less still with perceptual sensations, and far the least with somatic sensations. Strikingly, our sense that consciousness is intrinsic to mental states increases accordingly. The less useful it is to know about a particular kind of mental state even when the person is unaware of it, the more compelling is the intuition that that kind of mental state must be conscious.[10]

---

[10] Rosenthal (1986), 348.

We might also say: the less useful it is to know about a particular kind of mental state even when the person is unaware of it, the more compelling the intuition that the person can't be wrong about whether she is in that state.

Thus, Schiffer's worry needn't worry us. We're uninterested in "intense pains" which aren't conscious by virtue of our mechanisms being out of whack, because we wouldn't be inclined to count a state as <u>being</u> an intense pain unless we firmly believed we were in them. Likewise, we're <u>very</u> interested in appearances of being in pain--indeed, so interested as to count such appearances as pains whether or not they're accompanied by states which, for theoretical or common-sense reasons, we list under the extension of 'pain'. So long as we note that being inclined or disinclined to count something as a pain on account of our interests is not in itself a conclusive reason for thinking a state is or is not a pain, we can accept that, for practical purposes, some states which are pains won't count as pains, and some states which aren't pains will.

This, I think, is the force of the transcendental account. This kind of account asserts that there are

practical and conceptual constraints on ascriptions, and we needn't deny that these constraints exist. We must merely keep in mind that, for theoretical purposes, they're defeasible.

Therefore, I suggest that our non-empirical authority has a basis in the reliability of the mechanisms responsible for conscious thought and sensation; but I allow that practical and conceptual considerations make this authority seem more absolute than it is in reality.

## 2. Final thoughts

If I'm right about IMA, then there isn't really any problem about externalism and self-knowledge, and there never was any problem. I don't see myself as having solved a paradox, but merely as having dispelled the appearance of paradox.

Still, there is some use in this. There are complex epistemic and semantic issues at the heart of the apparent paradox. Externalism comes in various versions, not all of them plausible. Privileged access admits of two interpretations, only one of them plausible in light of externalistic considerations. Non-empiricality comes in

four flavors, two of which arose through examination of the effect externalism has on the concept of empirical defeasibility. If the discussion here has helped clarify our understanding of these notions, then merely dispelling the appearance of paradox has been well worth the effort.

# BIBLIOGRAPHY

Almog, Joseph (1981). 'Dthis and Dthat: Indexicality Goes Beyond That', Philosophical Studies 39, 347-81.

Alston, William (1971). 'Varieties of Privileged Access', American Philosophical Quarterly 8, 223-41.

Armstrong, David (1969). A Materialist Theory of Mind (London, Routledge and Kegan Paul).

Bach, Kent (1988). 'Burge's New Thought Experiment: Back to the Drawing Room', Journal of Philosophy 85, 88-97.

Boghossian, Paul (1989). 'Content and Self-Knowledge', Philosophical Topics 17, 5-26.

Bonjour, Laurence (1991). 'Is Thought a Symbolic Process?' Synthese 89, 331-52.

Brown, Jessica (1995). 'The Incompatibility of Externalism and Privileged Access', Analysis 54, 149-56.

Brueckner, Anthony (1986). 'Brains in a Vat,' Journal of Philosophy 83, 148-67.

----- (1990). 'Scepticism about Knowledge of Content', Mind 99, 447-51.

----- (1992). 'What an Anti-Individualist Knows A Priori', Analysis 52, 111-18.

----- (1994). 'Knowledge of Content and Knowledge of the World', Philosophical Review 103.

Burge, Tyler (1979). 'Individualism and the Mental', Midwest Studies in Philosophy IV, 73-121.

----- (1982). 'Other Bodies', in Woodfield (1982).

----- (1988). 'Individualism and Self-Knowledge,' Journal of Philosophy 85, 649-63. Reprinted in Cassam (1994).

Cassam, Qaussim (1994). Self-knowledge (Oxford: Oxford University Press).

Christensen, David (1993). 'Skeptical Problems, Semantical Solutions', Philosophy and Phenomenological Research 53, 301-21.

Church, Alonzo (1954). 'Intensional Isomorphism and Identity of Belief,' Philosophical Studies 5, 65-73.

Cohen, Stewart (1988). 'How to Be a Fallibilist', in James Tomberlin (ed.), Philosophical Perspectives 4 (Atascadero, Ridgeview).

Crimmins, Mark and John Perry (1989). 'The Prince and the Phone Booth: Reporting Puzzling Beliefs,' Journal of Philosophy 86, 685-711.

David, Marian (1991). 'Neither Mentioning 'Brains in a Vat', nor Mentioning Brains in a Vat will Prove we are not Brains in a Vat,' Philosophy and Phenomenological Research 51, 891-96.

Davidson, Donald (1984). 'First-Person Authority,' Dialectica 38, 101-11.

----- (1987). 'Knowing One's Own Mind', Proceedings and Addresses of the APA 60, 441-58. Reprinted in Cassam (1994).

Davies, Martin (1994). 'Externalism, Architecturalism, and Epistemic Warrant,' Eastern APA Symposium Lecture.

Dennett, Daniel (1991). Consciousness Explained (Boston: Little, Brown).

DeRose, Keith (1995). 'Solving the Skeptical Problem', Philosophical Review 104, 1-52.

Dretske, Fred (1970). 'Epistemic Operators', Journal of Philosophy 67, 1007-23.

----- (1981). Knowledge and the Flow of Information (Cambridge, MA: The MIT Press).

----- (1995). Naturalizing the Mind (Cambridge, MA: The MIT Press).

Dummett, Michael (1973). Frege: Philosophy of Language (London: Duckworth).

Elugardo, Reinaldo (1993). 'Burge on Content', Philosophy and Phenomenological Research 53, 367-84.

Ericsson, Anders and Herbert Simon (1984). Protocol Analysis (Cambridge, MA: The MIT Press).

Falvey, Kevin and Joseph Owens (1994). 'Externalism, Self-Knowledge, and Skepticism', Philosophical Review 103, 107-37.

Field, Hartry (1974). 'Quine and the Correspondence Theory', Philosophical Review 83, 200-28.

----- (1977). 'Logic, Meaning, and Conceptual Role', Journal of Philosophy 76, 379-409.

----- (forthcoming). 'The A Prioricity of Logic'.

Fodor, Jerry (1990). A Theory of Content (Cambridge, MA: The MIT Press).

----- (1994). The Elm and the Expert (Cambridge, MA: The MIT Press).

Forbes, Graeme (1987). 'The Indispensibility of Sinn', Philosophical Review 99, 535-63.

Frege, Gottlob (1952). 'On Sense and Reference', in Geach and Black (eds.), Translations from the Philosophical Writings of Gottlob Frege (Oxford, Basil Blackwell).

Greenwood, John (1991). 'Self-Knowledge: Looking in the Wrong Direction', Behavior and Philosophy 19, 35-47.

Heil, John (1988). 'Privileged Access', Mind 97, 238-51.

Kripke, Saul (1980). Naming and Necessity (Cambridge, MA, Harvard University Press).

LaPorte, Joseph (1996). 'Chemical Kind Term Reference and the Discovery of Essence', Noûs 30, 112-132.

Lewis, David (forthcoming). 'Elusive Knowledge'.

Loar, Brian (1992). 'Self-Interpretation and the Constitution of Reference', in James Tomberlin (ed.), Philosophical Perspectives 8 (Atascadero, Ridgeview).

Ludlow, Peter (1995). 'Externalism, Self-Knowledge, and the Prevalence of Slow Switching', Analysis 55, 45-49.

Lycan, William (1987). Consciousness (Cambridge, MA: The MIT Press).

Mates, Benson (1952). 'Synonymity,' in Leonard Linsky (ed.), Semantics and the Philosophy of Language (Urbana: University of Illinois Press, 1952).

McGinn, Colin (1986). Mental Content (Oxford: Basil Blackwell).

McKinsey, Michael (1987). 'A Priorism in the Philosophy of Language', Philosophical Studies 52, 1-34.

----- (1991). 'Anti-Individualism and Privileged Access', Analysis 51, 9-16.

-----. 'Accepting the Consequences of Anti-Individualism', Analysis 54, 124-28.

Millikan, Ruth Garrett. Language, Thought, and Other Biological Categories (Cambridge, MA: The MIT Press).

Nisbett, R.E. and T.D. Wilson (1977). 'Telling More Than We Can Know', Psychological Review 84, 231-59.

Nozick, Robert (1981). Philosophical Explanations (Cambridge, MA, Harvard University Press).

Pollock, John (1986). Contemporary Theories of Knowledge (Totowa, NJ: Rowman and Littlefield).

Putnam, Hilary (1975). 'The Meaning of Meaning', reprinted in Putnam's Philosophical Papers II (Cambridge, Eng.: Cambridge University Press).

----- (1981). Reason, Truth, and History (Cambridge, Eng: Cambridge University Press).

Raffman, Diana (1994). Comment on Davies (1994).

Rosenthal, David (1986). 'Two Concepts of Consciousness', Philosophical Studies 49, 329-59.

----- (1991). 'The Independence of Consciousness and Sensory Quality,' in Enrique Villanueva (ed.), Philosophical Issues 1 (Atascadero: Ridgeview).

Ryle, Gilbert (1949). The Concept of Mind (New York: Barnes and Noble).

Salmon, Nathan (1986). Frege's Puzzle (Cambridge, MA: The MIT Press).

Schiffer (1977). 'Naming and Knowing', in P. French, T. Uehling, and H. Wettstein (eds.), Midwest Studies in Philosophy II (Minneapolis: University of Minnesota Press).

----- (1987). Remnants of Meaning (Cambridge, MA: The MIT Press).

----- (forthcoming). 'Contextualist Solutions to Scepticism'.

Searle, John (1983). Intentionality (Cambridge, Cambridge University Press).

Shoemaker, Sydney (1990). 'First-Person Access,' Philosophical Perspectives 4 (Atascadero: Ridgeview).

----- (1993). 'Self-Knowledge and "Inner Sense" (The Royce Lectures), Philosophy and Phenomenological Research 54, 249-314.

----- (1994). 'Introspection', in Jonathan Dancy and Ernest Sosa eds., A Companion to Epistemology (Oxford: Basil Blackwell).

Stroud, Barry (1984). The Significance of Philosophical Skepticism (Oxford, Oxford University Press).

Tye, Michael (1994). Ten Problems of Consciousness (Cambridge, MA: The MIT Press).

Tymoczko, Thomas (1989). 'In Defence of Putnam's Brain's,' Philosophical Studies 57, 281-97.

Warfield, Ted (1992). 'Privileged Self-Knowledge and Externalism are Compatible', Analysis 52, 232-37.

----- (1995). 'Knowing the World and Knowing Our Minds,' Philosophy and Phenomenological Research 55, 525-45.

Williams, Michael (1991). Unnatural Doubts (Cambridge, Basil Blackwell).

Wilson, Mark (1982). 'Predicate Meets Property', Philosophical Review 91, 549-89.

Woodfield, Andrew (1982). Thought and Object: Essays on Intentionality (Oxford: Oxford University Press).