2003

Disarming the externalist threat to self-knowledge

Gabriel Guy Cate
Louisiana State University and Agricultural and Mechanical College

DISARMING THE EXTERNALIST THREAT TO SELF-KNOWLEDGE

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Arts

in

The Department of Philosophy and Religious Studies

by
Gabriel Guy Cate
B.A., Louisiana State University, 2001
May 2003

ACKNOWLEDGEMENTS

I would like to give thanks to my fellow graduate students, Jason L. Megill and Paul Jude Naquin, for letting me interrupt their own research with my incessant questions and aggravations; and to Jen O'Connor for being so supportive and understanding of all of us graduate students. She is the glue that binds the department!

I also would like to thank the following three people for serving as my committee members: John Baker, Jon Cogburn, and Husain Sarkar. I thank Dr. Baker for having a kind ear throughout this busy semester. I thank Dr. Cogburn for his unparalleled enthusiasm and willingness to help me get started on this project.

Finally, I want to give special thanks to Dr. Sarkar, whose overwhelming support, encouragement, and faith in me has been a driving force in shaping who I have become in the last several years. A person is of quite a rare breed to inspire as much strength and confidence in another as he has inspired in me.

TABLE OF CONTENTS

ABSTRACT

The purpose of this thesis is to examine various attempts to disarm the externalist threat to self-knowledge. That threat is engendered by a certain causal theory of meaning and reference, which suggests that empirical investigations may be required to know the contents of our own thoughts. It is claimed, then, that direct, non-inferential self-knowledge of our own mental states, is not possible if externalism is true. The leading compatibilist strategies that attempt to reconcile these apparently conflicting theses are explored and criticized. I conclude by offering what I take to be the essential features of a more successful compatibilist strategy.

CHAPTER 1 INTRODUCTION

How might *you* react if someone were to tell you that you do not know the content of your own thoughts or that you do not know what you mean by your sincere first-person assertions? I believe most people would find such a claim rather counter-intuitive since we normally think that we have authority over or privileged access to our own mental states. That is, it seems that, since our way of knowing our own thoughts is different, indeed superior, to others' way of knowing our thoughts, it is absurd for someone else to contend that we may not know what we think we know. Descartes certainly held this intuition because, as far as he was concerned, we know our thoughts infallibly through introspection alone. No one could possibly be in a better epistemic position with respect to one's thoughts than the thinking individual himself.

Though most of us are unwilling to endorse Descartes' claims about the infallible nature of introspective knowledge (most of us think that we can be, and often are, mistaken about the truth of our thoughts), we generally believe that we do have knowledge of our mental states. Of course, this knowledge can be threatened by a psychological phenomenon like self-deception, but such instances are hardly standard and do not threaten self-knowledge in general. What does appear to threaten the epistemic specialness of self-knowledge is the notion that the contents of our thoughts are determined, at least in part, by external, environmental factors. Let me explain.

When debating the issue of self-knowledge, philosophers of mind and language do not have in mind the Socratic notion of what it means to "know thy self"– it is not a matter of whether we know who we are. Rather, having self-knowledge is having knowledge of the contents of our own thoughts, of our own mental states in such a way that is, in principle, different from other people's way of knowing our own mental states. Intuitively, we believe that self-knowledge has some sort of special epistemic status: we know what we are thinking in a

direct, non-inferential way; others can know what we are thinking too, but only in an indirect, inferential way. This difference suggests that self-knowledge is in some sense *privileged*.

The infallibility of self-knowledge claims, however, is not the issue. The issue, rather, is that we might not even know what we *mean* when we express our own thoughts, at least not directly or non-inferentially. It is one thing to attack Descartes' view on the infallibility of self-knowledge claims; it is quite another to contend that direct, non-inferential self-knowledge is not possible. The strong position (represented by Descartes) has come to be known as *internalism*, or the idea that an individual's thoughts depend upon nothing but the individual who has them. Very few people hold this view today. Instead, many people who deny internalism in favor of *externalism*, the idea that the external environment *does* play a role in determining the contents of our thoughts, also deny the possibility of self-knowledge. But, as I will show, denying *internalism* does not entail denying *self-knowledge*.

Hilary Putnam was the first person to clearly show the *prima facie* threat externalism poses for self-knowledge. He argued that if words are used to describe the contents of an agent's thoughts, then it is natural to claim that the agent must know the meanings of those words, if he is to know the contents they describe. However, according to his causal theory of meaning and reference, the meanings of words depend upon the relations the items represented by those words bear to the external world. Thus, knowing the meanings of those words requires the agent to know the external referent of the words he uses to describe the contents of his thoughts. However, since knowledge of those external factors cannot be the result of introspective processes, self-knowledge cannot have the special epistemic status we intuitively attribute to it. Thus, self-knowledge is threatened, at least *prima facie*, by the externalist thesis.

This consequence of the causal theory of meaning and reference does not sit well with those of us who wish to maintain the notion that we do authoritatively know what we think. In fact, Putnam's denial of self-knowledge has spawned a fair number of compatibilist approaches for reconciling self-knowledge and externalism. Tyler Burge, for example, has suggested that the mere fact of the external determination of meaning does not entail a threat to privileged self-knowledge. He claims that even if the meanings of words or concepts are determined by external factors, it is not the case that the agent, who employs words and concepts in his thoughts, must know the external conditions that must obtain in order for him to know what he is thinking. Moreover, self-knowledge is actually one's knowledge of one's second-order thoughts, i.e. "I know that I believe *that p*." The external determination of *that p*, on Burge's view, does not threaten the agent's claim to know that he believes *that p*.

Donald Davidson, as well, contends that externalism is compatible with self-knowledge because denying self-knowledge precludes the possibility of successful communication. He claims that first-person authority must be preserved given the role it plays in the nature of interpretability. That is to say, if one is to successfully interpret another's sincere first-person assertions, the interpreter must assume that the agent being interpreted knows what his words mean. The speaker may misconstrue the meanings of the words he uses, but this possibility does not further suggest that he does not know what he *thinks* his words mean. Moreover, the speaker does not usually doubt what he thinks his words mean. Therefore, successful interpretation of another's assertions must require the interpreter to assume that the speaker knows the meanings of his words in a direct and authoritative manner. The first-person authoritative character of self-knowledge is thereby saved from the *prima facie* externalist threat.

According to Akeel Bilgrami, however, the compatibilist strategies suggested by Burge and Davidson do not free self-knowledge from *all* externalist threats. If we are to provide a complete reconciliation of externalism and self-knowledge, i.e. a compatibilist strategy that avoids all threats and not simply the *prima facie* threat, we must free ourselves from Putnam's specific externalist thesis. In other words, Putnam's causal theory of meaning and reference engenders a specific externalist thesis that cannot, in principle, be reconciled with self-knowledge. It is not the case, however, that all externalist theses are committed to Putnam's views on the external determination of meaning. Bilgrami suggests a new specific version of externalism that incorporates an agent's complete set of beliefs for the determination of meaning. He claims that such an externalism can account for direct, non-inferential self-knowledge.

Chapter 2 is devoted to developing and comparing Putnam's causal theory of meaning and reference with the traditional theory of meaning and explains in detail the specific account of externalism to which it leads. In the last section of that chapter, I pose an objection to Putnam's causal theory of meaning and reference in an effort to disarm the externalist threat to self-knowledge at the ground level. It seems to me that Putnam's view of the role the external element plays in the determination of meaning is *prima facie* false. I argue that Putnam suggests a much too stringent connection between meaning and reference.

Chapter 3 provides a detailed account of Burge's compatibilist strategy and explores three possible counter-arguments. First, according to Paul Boghossian, Burge's compatibilism can only account for a subset of instances of self-knowledge, i.e. the second-order judgments about first-order thoughts. He claims Burge has not provided the necessary tools for accounting for the *general phenomenon* of self-knowledge. Second, Sven Bernecker argues that the attitudinal component of self-knowledge is left unaccounted for on Burge's view. Though Burge

successfully demonstrates a compatibilist account of the first-order thought, he has not provided an account of the particular second-order judgment I make about my first-order thought, i.e. whether I *believe*, *judge*, *doubt*, etc. my first-order thought. Finally, I argue that if the content of the second-order judgment (the first-order thought) is determined externally, then the second-order judgment must also be externally determined. Moreover, if the second-order judgment is dependent on the external world for its content, then there does not appear to be room for the agent's first-person authority – knowledge of the second-order judgment is not privileged.

In Chapter 4 I discuss Davidson's appeal to the nature of interpretation for preserving self-knowledge. I then raise an objection to his compatibilist strategy on the grounds that his commitment to a mental-physical identity theory entails that all mental states are physical states, which are publicly observable. He must, therefore, be willing to grant that all mental states are publicly observable. The objection is based on the notion that privileged access suggests that only the agent in a mental state can know that he is in that state. Therefore, Davidson's view is not compatibilist since it has no self-knowledge component. In section 4.3, however, I argue that the objection rests on a mistaken understanding of self-knowledge. It is not the case that an agent's mental states are in principle knowable only to the agent in those states. Rather, the agent's way of knowing his own mental states is privileged in the sense of not being the same as that of an outside observer.

Finally, Chapter 5 addresses Bilgrami's attempted reconciliation of externalism and self-knowledge. Bilgrami claims that his specific version of externalism is based on an anti-foundationalist theory of meaning; that is, he suggests that, since the concepts an agents employs in sincere first-person utterances are mediated by that agent's aggregate set of beliefs, which will inevitably vary from agent to agent, there are no concepts that have analytic meanings. After

developing this specific externalist position, I pose two objections to Bilgrami's compatibilist strategy: first, I argue that one cannot endorse his externalism *and* claim to avoid all definitional meanings of concepts; and second, I suggest that his strategy is not as removed from Putnam's own compatibilist strategy as Bilgrami thinks it is. This is, of course, only a *very* brief sketch of my arguments. Let me turn now to the heart of the matter.

CHAPTER 2 HILARY PUTNAM

2.1 PUTNAM'S THEORY OF MEANING AND REFERENCE

The *prima facie* threat the externalist levies against self-knowledge is most often considered a direct result of the theory of meaning put forth by Putnam in "The Meaning of 'Meaning'." It is therefore important to carefully explore Putnam's theory of meaning and its implications if we are to truly understand how it is that some philosophers come to doubt whether we know what we think we know. With Putnam, I urge the reader to "kindly assume that *nothing* is clear in advance." Traditionally, two assumptions have been the basis for determining the meaning of 'meaning': (1) an individual's psychological state determines the meaning of a word, and (2) the intension of a word determines extension. Putnam claims that these two assumptions cannot be jointly satisfied and, as we will see, his theory of meaning rejects the first assumption outright while retaining a version of the second. A full explanation of the traditional theory of meaning is in order before we can tackle Putnam's criticism of it.

When we say that a word *means* something or other, we often have in mind the notion that the word denotes an 'instance,' or token, of something. In this case, by 'means' we mean *extension*. The extension of a word is the set of all things of which the word is true. Thus, to say what 'rabbit' means is to denote 'that which belongs to the set of all rabbits.' Some words, however, have more than one sense, which can make determining membership in a set, or extension, more or less 'fuzzy.' In such instances, we simply think of the word as having an invisible subscript referring to the particular sense being used at any particular time. Furthermore, the idea of *truth* in this definition of 'meaning' (in the sense of extension) is problematic in its own right. After all, no single understanding of 'truth' is agreed upon. So to think of 'meaning' simply in terms of a word's extension requires severe idealizations about the

limits of a set and the nature of truth. Thus, the *intension* of a word is introduced to the traditional theory of meaning in an effort to help clarify what 'meaning' means. *Intension* is most often (and misleadingly) understood as 'concept,' according to Putnam. So, in one sense, the word 'meaning' means *extension*, and in another sense it means *intension*; quite an ambiguity for a theory of meaning which the tradition of philosophical discourse has so firmly presumed!

There are two consequences of this ambiguity in the traditional theory of meaning. First, concepts are traditionally thought of as something *mental*; and by implication from the (sketchy) definition of intension (the notion that *meanings* are *concepts*), meanings must, then, be understood as mental entities. In spite of the fears of such philosophers as Frege and Carnap, who maintained that *meanings* are public property, so to speak, and that thinking of meanings as mental entities suggests that they could not be 'grasped' by more than one person or at different times, it seems that *meanings* are, at least partially, *mental*. In other words, even if Frege and Carnap had their way, and meanings are identified with some sort of Platonic, abstract entities rather than with concepts, the 'grasping' of such abstract entities is a mental, or psychological, act nevertheless. This conclusion is exactly what the first assumption of the traditional theory of meaning encompasses: knowing the meaning of word is simply a matter of being in a certain psychological state.

Second, the traditional theory holds that two terms can have the same *extension* and yet differ in their *intension*. For example, the terms 'creature with a heart' and 'creature with a kidney' share the same extension (e.g. humans are members of the set of 'creature with a heart' *and* of the set of 'creature with a kidney'), though they do not denote the same concept, or intension. Moreover, the traditional theory of meaning has taken it for granted that it is *impossible* for two terms to share the same *intension* and yet differ in their *extension*. Putnam

believes this assumed impossibility is a result of the ancient and medieval philosophers who assumed that the concept corresponding to a word "must *always* provide a necessary and sufficient condition for falling into the extension of a term." (Putnam, 219)[1] Furthermore, other traditional philosophers believed that the concept of a word provided a criterion for recognizing whether a particular item in the world was a member of a set, or belonged to the *extension*, of the word in question. Thus, the traditional theory of meaning rests on the following two assumptions:

(I) That knowing the meaning of a term is just a matter of being in a certain psychological state (in the sense of 'psychological state', in which states of memory and psychological dispositions are 'psychological states'; no one thought that knowing the meaning of a word was a continuous state of consciousness, of course).

(II) That the meaning of a term (in the sense of 'intension') determines its extension (in the sense that sameness of intension entails sameness of extension). (Putnam, 219)

Putnam argues that these two assumptions cannot be "jointly satisfied;" therefore, a new theory of meaning, one that rejects one or more of these two assumptions, is needed if there is to be any *meaningful* discussion of 'meaning.'

The reference to 'psychological states' in assumption (I) suggests a further, implicit assumption regarding the virtual non-role the external world traditionally plays in the determination of *meaning* – what Putnam calls the *assumption of methodological solipsism*: "the assumption that no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom that state is ascribed." (Putnam, 220) In other words, the first assumption of the traditional theory of meaning endorses what has come to be known as *internalism*, or the idea that knowledge of the meaning of a word (or, perhaps more usefully, knowledge of a thought) requires only introspective processes. That the individual's psychological state determines the meaning of a word (or thought) is a direct result of this

[1] Putnam, "The Meaning of 'Meaning'."

assumption (of methodological solipsism), since it presupposes nothing other than the individual himself. This assumption requires certain restrictions to be placed on what can and cannot count as a 'psychological state': those that methodological solipsism will allow are called 'psychological states in the *narrow* sense' and those that are not allowed are called 'psychological states in the *wide* sense.'

If we endorse the traditional theory of meaning, we find, by assumption (I), that if *A* and *B* are two terms with different extensions, *knowing the meaning of A* and *knowing the meaning of B* are two different psychological states. Further, by assumption (II), we know that *A* and *B* must have different *intensions* as well. Knowledge of the meaning of a word is not obtained simply by 'grasping its intension;' one must know *which* intension he is grasping if he is said to have knowledge of it.[2] For example, if I know the meaning of the word 'wheel,' presumably I can 'grasp the intension' of its German synonym '*Rad*;' however, I may not know the *meaning* of the word '*Rad*,' unless I know that it is the intension of the word '*Rad*' that I am 'grasping' (as opposed to 'grasping' the intension of 'wheel'). (Putnam, 221) Here we begin to see that for Putnam the psychological state of 'grasping the intension' (assumption I) is not enough for knowing the meaning of the word.

Furthermore, if I_1 and I_2 are different intensions of *A*, *knowing that* I_1 *is the meaning of A* and *knowing that* I_2 *is the meaning of A* are two different psychological states. It is therefore impossible, given the assumption of methodological solipsism, for there to be two possible worlds in which an individual is in the same psychological state if in one world he knows that I_1 is the meaning of *A*, and in the other world, he knows that I_2 is the meaning of *A*. In other words, for every possible world in which an individual is in a certain psychological state the necessary

[2] Bertrand Russell makes a similar point in *The Problems of Philosophy* (see esp. p. 58).

and sufficient condition for being in the extension of *A* is the same. So, if the psychological state determines the *intension*, and, by assumption (II), the *intension* of *A* is a necessary and sufficient condition for *A*'s *extension*, then the psychological state determines the *extension* of *A*. By the public nature of psychological states, two individuals *can* be in the same state at the same time; and if they are in the same psychological state, they *cannot* understand a word differently. That is, according to the traditional theory of meaning, the *extension* of a word cannot differ if the two individuals share the same *intension* of the word. This is exactly what Putnam wants to deny.

In an effort to show this possibility, Putnam offers several thought-experiments. Suppose there is a planet completely identical with Earth, call it Twin Earth, in which there are English speakers, there are mountains, trees, animals, etc. The *only* difference between Earth and Twin Earth is that the substance that fills Twin Earth's oceans and lakes, and falls from the sky when it rains, etc. is not H_2O, but rather has some other chemical composition, say, XYZ. The two substances are completely identical with respect to all of their observable (phenomenological) properties, such that the inhabitants of Earth and Twin Earth use 'water' in the same manner: on both worlds 'water' is used for drinking, cleaning, cooling, etc. What is more, the English speakers on Twin Earth also denote this substance with the word 'water' (or, as I will denote it, water_{TE}). The difference between water_{E} (what we on Earth call *our* substance) and water_{TE} can, in principle, be discovered once the chemical compositions of the two substances are compared. Therefore, in this case, the word 'water' has two meanings (in the sense of extension): water_{E} has the extension of H_2O; water_{TE} has the extension of XYZ, and yet the *intension* of the word 'water' is the same for both Earthian and Twin Earthian speakers.

Now, suppose that it is some time before the chemical compositions of both water_{E} and water_{TE} are known (also, assume that the growth of scientific knowledge on Twin Earth is

parallel with that on Earth), say, 1750. Putnam's critical point can be summarized in the following statement: regardless of whether the speakers on either Earth or Twin Earth are aware of the extension of their word 'water', the extension of 'water' is the same. The set of all things that are true of a term is the same independent of our recognition that a certain thing is a member of the set or not. So, in this case, the English speaker on Earth ($Oscar_1$) and the English speaker on Twin Earth ($Oscar_2$) understood their term 'water' differently before they knew the chemical composition of 'water.' Therefore, their psychological states were different in 1750 from their psychological states in, say, 1950. In this scenario, then, it is possible that the psychological state of the speaker is not the sole determining factor of the meaning (in the sense of *extension*) of the word 'water'.

It might be objected that we are not compelled to accept the idea that 'water' has the same extension in 1750 that it does in 1950.[3] After all, knowing the chemical composition of water will allow one to demarcate 'pure water' from 'polluted water.' So it seems reasonable that the extension of the word 'water' in 1750 might be different from its extension in 1950, a time when we are thought to be very good at demarcating 'clean' water from 'dirty' water. That is to say, in 1750 no one could distinguish pure H_2O from impure water, e.g. from the water in the Mississippi River. Putnam responds to this line of objection by invoking the notion of natural-kind terms.

When we say, 'this liquid is water,' we are giving an 'ostensive definition' of the word 'water.' That is to say, our claim presupposes the empirical fact that the substance we are referring to bears a certain sameness relation ($same_L$) to most of the other 'stuff' in the world that we call 'water.' Obviously, if this presupposition is false and the substance does not bear that

[3] See section 2.3 for a full account of such an objection.

same$_L$ relation, we have reason to doubt the 'ostensive definition' that was given. However, the presupposition need not be infallible if we are to endorse the 'ostensive definition.' In other words, the same$_L$ relation is *theoretical* and though it may be shown that the empirical presupposition upon which the same$_L$ relation was based is false, this does *not* mean that the extension of the word 'water' changes when the same$_L$ relation does not hold. Indeed, the psychological state of Oscar$_1$ in 1750 is different from his psychological state in 1800 (assuming, of course, that the chemical composition of water is discovered sometime between 1750 and 1800, and, moreover, that this information is imparted to Oscar$_1$ by 1800). This difference in psychological state, however, has no bearing on the extension of the word 'water.' Thus, according to Putnam, we are justified in holding that the extension of 'water' is no different in 1750 than it is in 1950 (or at any point in time, for that matter).

Putnam makes a similar point with a variation of this Twin Earth thought-experiment. Suppose, he says, that molybdenum and aluminum are like H_2O and XYZ, their phenomenological properties are identical. The only difference between them is their chemical composition. On Earth, aluminum is used to make pots and pans; on Twin Earth molybdenum is so used. Further, in the Twin Earth idiolect 'aluminum' refers to *molybdenum*, so that when Twin Earthian speakers say 'aluminum' (aluminum$_{TE}$ = molybdenum$_E$, if you will) they *mean* molybdenum (in *our* idiolect). Now suppose that though the standard Earthian and Twin Earthian speakers are unable to distinguish the difference between (what we call) *aluminum* and *molybdenum*, Earthian and Twin Earthian metallurgists *can* make such a distinction. The difference between this Twin Earth thought-experiment and the previous thought-experiment is subtle, but very important. Whereas *no one* on Earth or Twin Earth in 1750 could have discriminated between water$_E$ and water$_{TE}$ (H_2O and XYZ), only those non-metallurgists (only a

portion, large as it may be, of standard English speakers on Earth and Twin Earth) are unable to discriminate between aluminum_E and aluminum_{TE}.

The point Putnam is trying to make here is that there is a *division of linguistic labor* at work in our linguistic community such that not everyone is held responsible for being able to *recognize* whether *x* belongs in the extension of *X*. It is sufficient that *someone* possesses that way of recognizing for the *community* to be said to possess it. Thus, it is possible for me to *know that water is* H_2O without having to perform some sort of chemical analysis of a liquid I believe to be *water*. Moreover, this thought-experiment demonstrates that Oscar_1 and Oscar_2 are in the exact same psychological state when they use the word 'aluminum,' yet the extension of the word 'aluminum' is certainly different on Twin Earth from its extension on Earth. Driving it home one more time, then, we see that it is *possible* for two people to be in the exact same psychological state when they use a word, say, 'aluminum,' and the extension of that word is different in their respective environments. It must, therefore, be false that psychological states are solely responsible for determining the extension of a word (methodological solipsism is false.) As Putnam tells us, "cut the pie anyway you like, 'meanings' just ain't in the *head*!" (Putnam, 227) Let us turn now to assumption (II) to see if we can salvage anything from the traditional theory of meaning.

There are two ways one can tell someone else what he means by a natural-kind term (such as 'water,' 'tiger,' or 'lemon'): (1) he can give an 'ostensive definition,' or (2) he can give a description of the term. (Putnam, 231) Recall the Twin Earth (1750) thought-experiment: W_1 refers to Earth, where water_E is H_2O; W_2 refers to Twin Earth, where water_{TE} is XYZ. Now suppose that it is logically possible for an individual (Oscar_1) to have a *Doppelganger* (a

completely identical twin, $Oscar_2$). In W_1 the liquid $Oscar_1$ refers to as 'this (liquid) in the glass' is H_2O, and in W_2 $Oscar_2$ refers to XYZ. Two theories of meaning arise from this situation:

(1) 'Water' is world-relative, but constant in meaning; that is, 'water' means the same in W_1 and W_2 if, and only if, $water_E$ is H_2O and $water_{TE}$ is XYZ, or, more explicitly,

(1') (For every W) (For every *x* in W) (*x* is water iff *x* bears $same_L$ to the entity referred to as 'this' in W)

Or,

(2) 'Water' does not have the same meaning in W_1 and W_2; 'water' is H_2O in all worlds, thus $water_{TE}$ is *not* water, or more explicitly,

(2') (For every W) (For every *x* in W) (*x* is water iff *x* bears $same_L$ to the entity referred to as 'this' *in the actual world* W_1) (Putnam, 231)

The Twin Earth thought-experiment suggests that by 'water' we mean (2'). Saul Kripke calls this definition of a word *rigid designation*. "If we extend the notion of rigidity to substance names," the word 'water' is a rigid designator in the theory of meaning (2') since the word "refers to the same individual in every possible world in which the designator designates." (Putnam, 231) The consequence this theory of meaning has on the theory of necessary truth is that once the chemical properties of water are known to be H_2O, there are no logically possible worlds in which water is not H_2O. In other words (Kripke's in fact), the statement, 'water is H_2O' is 'metaphysically necessary.' Furthermore (adding more fuel to the externalist fire), Putnam claims, "human intuition has no privileged access to metaphysical necessity." (Putnam, 233) That is to say, though we may think that 'water' has the same meaning on Earth and Twin Earth, since both $Oscar_1$ and $Oscar_2$ have the same 'operational definition' of 'water,' 'water' in fact has only one meaning, which is satisfied only if the substance in question bears the $same_L$ relation to the stuff in the *actual* world (W_1, or Earth).

The point of all this is to say that natural-kind terms have a certain *rigidity*, or, as Putnam says, *indexicality*. Words like 'now,' 'this,' 'here,' and, in particular, 'I,' have been recognized

as '*token-reflexive*' – the 'objects' they denote bear a $same_L$ relation to other stuff in the environment. However, the notion that 'intension determines extension' has not been suggested for such indexicals. But, if my twin on Twin Earth has the thought 'I have a headache' whenever I have the thought 'I have a headache,' then surely the 'I' he uses has a different extension than the 'I' I use, even though our intensions are the same (his concept of *himself* is not different from the concept I have of *myself*).[4] If it is true that natural-kind terms are indexical, then the meaning (extension) of the natural-kind term 'water' cannot be determined by its intension.[5] So, assumption (II) is problematical if we assume (I).

What have we learned so far? Primarily, Putnam has given us two reasons to think that extension is not fixed by a 'concept' an individual might have in his head. First, intension has a social dimension due to the division of linguistic labor. Second, intension (in this new sense) exhibits a kind of indexicality or rigidity. We are now faced with two paths for understanding the meaning of 'meaning.' Either, ignore these two reasons and retain assumption (I), which identifies meaning (in the sense of intension) with concept; or, reject assumption (I) and identify 'meaning' with an ordered-pair of entities: the social dimension of the meaning of a word and the *extension* of the word.[6] If we take the first path, we must conclude that a word can have the same meaning (in the sense of intension) on Earth and Twin Earth, and yet differ in extension. As the above paragraph explains, this is fine for 'absolutely' indexical words like 'I,' but it is not clear that this result is appropriate for other, (less indexical?) natural-kind words like 'water.' Putnam takes the discussion about the metaphysical necessity that water is H_2O in all possible

[4] My twin's concept of 'self' and my concept of 'self' are the same, or identical, in the sense in which two neckties can be the same or identical – identity here does not mean 'numerical identity.'
[5] I do not mean to suggest that 'I' is a natural-kind term; rather, natural-kind terms display a kind of indexicality similar to that of *indexicals* like 'I'.
[6] More about this ordered-pair will be explained later in the chapter.

worlds as evidence against the first route since it requires us to have invisible subscripts for innumerable meanings of the same word.

If we take the second path, we must abandon the idea that a difference in the meaning of a word both my twin and I use *entails* some difference in our psychological states. Once again, the Twin Earth thought-experiment leads us to this conclusion. It was shown, in the case of aluminum_{E} and aluminum_{TE}, that my twin and I are both linguistically competent and yet mean different things by the word 'aluminum.' If we take this second path, we are able to see that the problem of determining a theory of meaning is actually two problems. First, a theory of meaning must account for the determination of extension. Putnam suggests that extension is socially determined (i.e. thanks to the division of linguistic labor) and defers the issue to socio-linguistics. The second problem is the problem of determining how to hold speakers accountable, so to speak. That is, it is a problem of describing the linguistic competence of the speaker.

Individuals have to have "some particular ideas and skills in connection" with the actual world if they are to play a part in the linguistic division of labor. (Putnam, 246) We cannot simply let people use words however they want; to do so would be to lose all hope of any meaningful communication. First, speakers in a linguistic community must agree upon certain grammatical rules most commonly understood in terms of 'syntactic markers' like 'noun,' 'adjective,' 'adverb,' etc. Second, speakers must also have a common vocabulary of 'semantic markers' like 'water,' 'aluminum,' etc. Finally, Putnam suggests that stereotypes are used in an effort to help others 'acquire' new words so that 'significant communication' can take place. One can *acquire* a word without *knowing* it (in the sense of knowing its extension). Stereotypes are idealizations of the extension of a word. Therefore, it is, of course, possible for stereotypes

to incorrectly describe the extension of a word. This fact, however, should not deter us from using them. Stereotypes merely serve 'operationally;' that is, as convenient idealizations, they facilitate acquiring new words in order to successfully communicate. In a sense, it is linguistically obligatory to use stereotypes when helping others acquire new words; otherwise, no discussion could begin.

We are now at a point where we can clearly lay out Putnam's theory of meaning. He defines 'meaning' "by specifying a normal form (or, rather, a *type* of normal form) for the description of 'meaning.' This normal form description is a finite sequence (vector) of components including the following:

1) The syntactic markers that apply to the word, e.g. 'noun'
2) The semantic markers that apply to the word, e.g. 'animal', 'period of time'
3) A description of the additional features of the stereotype, if any
4) A description of the extension. (Putnam, 269)

Components 1) – 3), when taken together (as one component of the ordered-pair), determine the linguistic competence of the speaker. The extension of the word (the other component of the ordered-pair) is independent of the speaker (a direct result of the Twin Earth thought-experiments). What this amounts to is saying that two equivalent descriptions (independent of the extension) accurately describe the meaning of a word if they are coextensive and describe a set that is, in fact, the extension of the word in question. Thus, assumption (I) is rejected on the grounds that psychological states do not determine extension, either directly or indirectly (e.g. via determining intension, which in turn, determines extension); and assumption (II) is retained with the understanding that intension is not determined by the psychological state of the speaker.

2.2 PUTNAM'S EXTERNALISM

Now we are in a position to answer the following question: how do the externalists use Putnam's theory of meaning as the foundation for their claims against self-knowledge? Minimally (and

probably most unhelpfully), we can answer this question by pointing to the fact that Putnam's theory completely abandons the assumption of methodological solipsism. Recall that this assumption claims that "no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom that state is ascribed." In other words, the internalist claims that only introspective processes are required for knowing the content of our own minds. Some externalists understand this claim as implying something much stronger than what the internalist intended. Such externalists assume that the internalist position with respect to self-knowledge is only possible if the assumption of methodological solipsism is true. It is very easy, then, for the externalist to hitch their cart to Putnam's theory of meaning since it flatly denies that assumption. Since externalism is most generally construed as the denial of this assumption, Putnam's theory of meaning is, generally, an externalist thesis. However, Putnam's externalism runs much deeper than merely denying the assumption of methodological solipsism.

Let us look now at how Putnam's theory of meaning specifically influences the externalist threat against self-knowledge. It has been widely argued (though not widely accepted) that if externalism is true, individuals may not know the content of their own thoughts. This claim is partly a result of the claims Putnam has made regarding the necessary and sufficient conditions for knowing the meaning of the words individuals use when describing their thoughts and beliefs. As Putnam's Twin Earth thought-experiments suggest, any speaker who is unable to determine whether the substance bears the $same_L$ relation, and is thus unaware of the proper extension of a word, say, 'water,' does not know what he thinks he knows about 'water.' Thus, when my twin expresses his thought that 'water is wet' he cannot know the content of his thought unless he understands the meaning of the word 'water' in the proper Putnamian sense of knowing the extension of 'water,' linguistic competence aside. Moreover, I cannot know the

content of my thought 'that water is wet' unless I know that the substance I am referring to in my thought is H_2O; luckily, scientists can perform the proper tests for me!

Not only must we claim that there are things in the world other than the individual (as externalism is very generally understood), now, thanks to Putnam's theory of meaning, we must also know the *extension* of the words expressed in our thoughts if we are to know the content of our thoughts. But if this is true, then Putnam and his externalist followers are committed, for example, to the idea that the current scientific definition of water (that it is H_2O) is the ultimate arbitrator in demarcating the bounds of water's extension. Furthermore, they are also committed to saying that even the most informed scientist in 1750 did not know what he meant when he made claims about 'water.'

The meaning of 'self-knowledge,' then, from the perspective of the externalist thesis influenced by Putnam, can be simply stated as follows: to have self-knowledge is more than having merely performed introspective processes alone; it is to have become aware of the full extension of the content of one's thoughts or beliefs. The content of one's thoughts and beliefs is simply the 'that-clause' in a statement such as "I think *that this is water*." In order to know 'that this is water,' one must know what 'water' means. According to the Putnamian externalist, knowing what the word 'water' means requires recognizing (through empirical methods) that it bears a sameness relation ($same_L$) to other things in the external world that are water.

Moreover, the necessary and sufficient conditions for bearing the $same_L$ relation are both that the substance I think is water *and* the other substances in my environment which I (and others) also think are water have the chemical composition H_2O. Therefore, if the particular substance I am currently having a thought about is H_2O, I do know the content of my thought. Otherwise, I do not know the content of my thought. And according to Putnam's theory of

meaning and the externalist thesis it influences, the only way to know that *this water* is H_2O is to either perform external, empirical investigations by myself or have the scientific experts in my community perform them.

In short, then, Putnam's theory of meaning requires strict guidelines for knowing the meaning of the content of a thought. The externalist position is based on this theory of meaning because it provides reasons for rejecting the assumption of methodological solipsism, which implies that the individual who has the thought is solely responsible for the meaning of the content of a thought. The stronger externalist position that the meaning of the content of one's thought is a matter of metaphysical necessity and knowledge of that metaphysical necessity can only be gained through external procedures is certainly influenced by Putnam's theory of meaning. The externalist position makes knowledge of any sort rather limited. Self-knowledge appears to be defunct outright. Further, there appears to be no reason to claim that thoughts are metaphysically necessary. So, if it is not the case that we are solely responsible for the meaning of the content of our thoughts *and* the only knowledge that counts is the knowledge of metaphysical necessity, self-knowledge is not possible. The best we can hope for is that someone else (an expert?) can tell us what we are thinking.

2.3 AN OBJECTION TO PUTNAM'S THEORY OF MEANING AND REFERENCE

One strategy for disarming the externalist threat to self-knowledge is, of course, to undercut the theory of meaning on which the externalist position rests. That is to say, a successful argument against Putnam's theory of meaning and reference would prevent the externalist threat from the get go.[7] So, how might such an argument against Putnam run? I believe that we can accept much of what Putnam has put forth in his theory of meaning (with a few caveats or

[7] Husain Sarkar presented the following objection to me in "A Rough Sketch of Two Counter-Arguments", February 25, 2003. I find it to be a rather devastating criticism of Putnam's theory of meaning.

modifications) and yet draw strikingly different conclusions about the meaning of 'meaning' and 'reference.' More importantly for our discussion, we can agree with many of Putnam's claims and yet disagree with the necessary and sufficient conditions for self-knowledge.

Recall Putnam's thought-experiment: Earth and Twin Earth are exactly the same in all respects (i.e. on both planets there are communities of English speakers, scientific growth is the same on each planet, etc.) except for the fact that the substance that fills Earth's lakes and oceans, falls from the sky as rain, is used for drinking, etc. has the chemical composition H_2O; whereas on Twin Earth this substance has the chemical composition XYZ. All phenomenologically observable properties of 'water' are the same on Earth and Twin Earth.[8] When $Oscar_1$ speaks of 'water' and $Oscar_2$ speaks of 'water' both individuals are in the same psychological state, yet they mean something different by the word 'water;' they refer to different substances, i.e. $water_E$ (H_2O) and $water_{TE}$ (XYZ), respectively. The extension of the word 'water,' according to Putnam, is different on Earth than it is on Twin Earth. His ultimate point, then, is that sameness of meaning necessitates sameness of reference. Since $Oscar_1$ and $Oscar_2$ refer to different substances when making utterances about 'water,' the word 'water' must mean something different when spoken from the mouths of $Oscar_1$ and $Oscar_2$.

Putnam's theory of meaning is quite dependent upon the notion that words have a certain indexicality, or rigidity, that is entailed by the $same_L$ relation. For *the liquid in this glass* to be water, it must bear the $same_L$ relation to $water_E$; it must have the chemical composition H_2O. To know that *the liquid in this glass* is water, one must know that it bears the $same_L$ relation, either by one's own experiments or through the division of linguistic labor, to the other stuff in the world. However, Putnam is all too eager to set the necessary and sufficient conditions of the

[8] I say 'phenomenologically observable' in the sense of being easily apparent to the senses; contrast these with other observable properties such as the microstructural properties.

same$_L$ relation. Putnam's determination of the necessary and sufficient conditions for bearing the sameness relation is simply arbitrary. The growth of scientific knowledge is the same on Earth and Twin Earth, *ex hypothesi*, so why is it that *our* water's chemical composition is the necessary and sufficient condition for bearing the same$_L$ relation if *their* water's chemical composition is discovered at the exact same time? Suppose Twin Earth scientists had performed chemical analysis of water$_{TE}$ *before* Earth's scientists performed chemical analysis of water$_E$. According to Putnam's theory of meaning, 'water' would *mean* XYZ; for the liquid in the glass to be water it must bear the same$_L$ relation defined by the chemical composition XYZ.

I should remark here that the heart of the following objection to Putnam's theory of meaning is its attack on the central claim that two individuals who are in the same psychological state can refer to different items (e.g. water$_E$ and water$_{TE}$, or aluminum$_E$ and aluminum$_{TE}$) with the same word. If they do not know the meaning of the words used to express their thoughts, they cannot know the content of their thoughts and so cannot have self-knowledge. So, if my twin and I have the same thought, "Water is wet," my twin would be thinking of water$_{TE}$ and I would be thinking of water$_E$. I can know that I am, in fact, referring to H_2O because the experts in my community have told me that 'water' means H_2O (and, of course, because I assume the only water-like substance in my community is water, or H_2O).

Putnam must claim, then, that it is possible for experts to be in the same psychological state and yet distinguish water$_E$ from water$_{TE}$. He does not, however, argue for this conclusion though it follows directly from the case of Oscar$_1$ and Oscar$_2$ being in the same psychological states and yet referring to two different substances. Suppose that an expert in contact with the observable properties of 'aluminum' is in psychological state S_1; if an expert knows the chemical

composition of aluminum$_E$ he is in psychological state S_2; and if an expert knows the chemical composition of aluminum$_{TE}$ he is in psychological state S_3.

If an expert knows the meaning of the word 'aluminum' as expressed in both Earthian and Twin Earthian idiolects, he must know the chemical composition of both aluminum$_E$ and aluminum$_{TE}$ and he must be able to tell the difference between the two simply by inspecting their observable properties. Furthermore, as the Oscar cases imply, the expert could be in the same psychological state, S_1, when confronted with aluminum$_E$ as he is when confronted with aluminum$_{TE}$ (molybdenum), since their phenomenologically observable properties are the same. Since, *ex hypothesi*, simply observing the two objects will not yield knowledge of the chemical composition of either substance, the expert must be in psychological state S_2 when he knows the meaning of aluminum$_E$ and in psychological state S_3 when he knows the meaning of aluminum$_{TE}$. Putnam cannot, and does not, argue for this conclusion because he wants to preserve the possibility of being in one psychological state and yet referring to different objects. Clearly, though, the expert cannot be in the same psychological state when referring to aluminum$_E$ and aluminum$_{TE}$. We can see, then, why Putnam does not argue the case of the experts.

All arbitrariness of the baptism of a word's meaning aside and despite this serious hole in Putnam's argument, according to his theory of meaning and reference, the only substance to which the word 'water' can refer must be a substance with the chemical composition H_2O. However, it is not entirely clear that we must endorse this conclusion. Take, for example, other English words that denote natural objects like 'leaf,' 'animal,' 'tree,' etc. compared with words that refer to man-made objects like 'chair,' 'table,' 'cup,' etc. The meanings of the words in the former set are not undermined by the fact that particular leaves (animals and trees) have

particular biological or chemical structures. That is, suppose I cannot tell the difference between the leaves of a maple tree and the leaves of a magnolia tree. The meaning of my word 'leaf' is no different when I say, "This leaf is beautiful," regardless of whether I am referring to a maple leaf or a magnolia leaf. Likewise, our use of the words in the latter set is equally pragmatic. The meaning of my word 'chair' in my utterance, "This chair is comfortable," does not change when referring to different chairs though the reference, obviously, does change. Putnam's insistence that meaning, in the sense of *extension*, is determined or fixed ultimately by an object's microstructural properties seems plainly false.

For Putnam (or Kripke or anyone else) to respond by simply saying, "But the substances *are* different, so how can the extensions not also be different?" is to beg the question. That is, the response presupposes that the ultimate determination of a word's extension *is* the microstructural property of the object to which the word refers. If we consider pragmatic concerns, difference in microstructural properties has no bearing on the meaning of a word. Suppose the following: (1) Throughout the histories of Earth and Twin Earth no one has taken into account the chemical structure of 'water' in determining the meaning of the word 'water,' they have only considered the phenomenologically observable properties; (2) speakers from Earth and Twin Earth can communicate with each other quite successfully (they both speak English); (3) both communities use 'water' for washing, drinking, irrigating, etc.; and (4) water_{E} and water_{TE} blend together just as two buckets of water_{E} would blend. Now, suppose the chemical composition of both water_{E} and water_{TE} were discovered. I see no reason to suspect that speakers in either community would *use* 'water' differently, both with respect to mentioning the word 'water' and with respect washing and irrigating with it, drinking it, etc. The additional information about the substance in each community has no effect on how 'water' is used.

We need not abandon all of Putnam's theory of meaning, however, in order to have a useful causal theory of meaning and reference. We could, with a simple caveat, endorse his notion of 'indexicality' or Kripke's 'rigid designation.' That is, without making the microstructural property of a substance the ultimate criterion for the 'baptism' of the word, we can make great use of the idea of the indexicality of words. If we take a hard line pragmatic approach, we can simply designate a term as referring to an object comprised of a certain set of phenomenologically observable properties, $O_1, O_2, O_3, O_4, \ldots, O_n$. Given this theory of meaning and reference, water_E and water_{TE} would be in the same extension, denoted simply by the term 'water.' This approach to understanding the meaning of 'meaning' will allow us to be as general or as specific as we care. In fact, this view is more in line with everyday experience. For example, the word 'leaf' refers to the set of all objects with phenomenologically observable properties, $O_1, O_2, O_3, O_4, \ldots, O_n$. The meaning of the word 'leaf' is very general. We make our concepts more specific by adding a qualifying adjective before the term, i.e. maple leaf. Certainly, the phenomenologically observable properties of a maple leaf will include all those properties necessary for generally being a leaf, but it will have another set of observable properties in addition. This set could include those properties that make it a maple leaf as opposed to a magnolia leaf; certainly one of those additional properties could be its specific microstructural properties, but this does not suggest that the microstructural properties are the ultimate external determinant of the meaning of the term 'maple leaf.' The microstructural properties are simply used to classify types of leaves, but they do not confer meaning upon the terms.

This brings us to an interesting point about the possibility of knowing all the properties of an object. So far I have been referring to the 'phenomenologically observable properties' with

the aim of distinguishing those observable properties that are easier to observe and, therefore, more widely known, from those properties that are less easy to observe, i.e. chemical structures, and so less widely known. But suppose the ultimate microstructural properties of 'water' are, in principle, unknowable given certain limitations of the mental capacities of the residents of both Earth and Twin Earth. Further, both communities have discovered all the phenomenologically observable properties it is, in principle, possible for the beings in the two communities to discover. If we follow Putnam, we must claim that neither my twin nor I know what we mean by 'water' since the criterion for knowing the meaning of 'water' is unknowable. Our tongues will be are tied; that is, we can never have successful communication if the ultimate criterion for determining the extension of a term, its ultimate microstructural composition, is, in principle, unknowable to us. Since we do communicate successfully, and often times efficiently, there must be something other than the ultimate microstructural properties of objects that determines the meanings of our terms and allows us to share a common language. There are therefore no pragmatic grounds for endorsing Putnam's theory of meaning and reference.

It seems, then, that Putnam's normal form definition of 'meaning' is too restrictive. His claim that both components of the ordered-pair are necessary and (together) sufficient for knowing the meaning of a word requires too much of the standard speaker. We can endorse the first component of the ordered-pair (linguistic competence) as being necessary for knowing the meaning of a word, without endorsing the second (extension). That is, being a linguistically competent speaker (alone) is sufficient for knowing the meaning of a word. Putnam makes severe idealizations about the limits of a word's extension when he claims that the truth about water is that it is H_2O. The nature of scientific progress is that we are always improving and approaching the truth, but that there is always more to be known. Putnam takes the perspective

of an all-knowing agent who either confers or does not confer knowledge to individuals. Given the nature of scientific progress, Putnam is not entitled to such a perspective.

The objection presented in this section provides at least one undercutting reason against Putnam's theory of meaning: it restricts too much of our knowledge. If we rid ourselves from the constraints of this theory of meaning we are free to reject the externalist's skeptical position with regard to self-knowledge it has so widely influenced. We can accept many of Putnam's claims without concluding that our self-knowledge is quite limited. In the following two chapters I will develop the, not entirely unproblematic, compatibilist positions of Tyler Burge and Donald Davidson and possible objections to their claims with respect to reconciling externalism and self-knowledge.

CHAPTER 3 TYLER BURGE

3.1 BASIC SELF-KNOWLEDGE

In "Individualism and Self-Knowledge," Tyler Burge presents what has become the most widely accepted compatibilist position with respect to privileged access to our own thoughts and the externalist thesis. The problem all compatibilists face is the apparent incompatibility of privileged access to one's thoughts and the external individuation of those thoughts. As Putnam's thought-experiments have shown, the content of one's thoughts is dependent upon one's external environment. Burge contends, however, that epistemic reliabilism grants that even though certain external conditions must obtain for certain thoughts to be what they are, knowing that the thought has occurred does not depend upon knowing that those external conditions actually obtain.

The heart of Burge's compatibilism is that though the content of a first-order thought depends on the environment, the special epistemic status of second-order judgments about first-order content is not threatened. For example, I have a first-order thought: "I think that writing requires concentration," which is accompanied by a coincidental second-order judgment: "I judge: I think that writing requires concentration." The content of the first-order thought is embedded or contained in the second-order judgment. Thus, the second-order judgment is not only self-referential, it is self-verifying since "making the judgment itself makes it true" – no empirical investigation is needed to know the second-order judgment.

Burge adopts a restricted Cartesian conception of what it means to know one's thoughts directly and authoritatively. He tells us that even though Descartes "tended to overrate the power of authoritative self-knowledge … [Descartes] was right to be impressed with the directness and

certainty of some of our self-knowledge." (Burge, 468)[1] Those instances of self-knowledge Burge finds undeniably direct and authoritative are the self-verifying second-order judgments. He calls these judgments *basic self-knowledge*.

The restricted Cartesian view is simply that first-order *thoughts*, or, if you prefer, *cogito-like* thoughts, are the most probable candidates for being self-verifying. Burge does not extend basic self-knowledge to knowledge of beliefs, desires, expectations, etc. because he does not agree with Descartes' claim that introspective knowledge allows us to 'cut off' the external individuating conditions that determine the content of such knowledge claims. That is, Descartes argued that, since we can think thoughts while doubting the existence of the external world, our thoughts (which Descartes generally construed as encompassing beliefs, desires, etc) are not dependent upon the external world for their existence or their content. From this, Descartes makes the further claim that we must be infallible with respect to these 'thoughts' since nothing external could deny their veridicality.

Externalism, very generally defined, denies both the possibility that thoughts are solely contingent upon the individual and the non-existence of the external world. Further, since the external world does exist, externalists are quick to dismiss Descartes' claim regarding the infallibility of our thoughts (since something external to the individual could deny the veridicality of some of our thoughts), and in so doing, they make the stronger claim that our thoughts cannot have any special epistemic status since they do indeed depend on the external world for their content. Burge, however, tells us that it is one thing to disagree with Descartes' conception of thought individuation as infallible; it is another to suggest that we do not know our thoughts non-empirically. Let us return to a Twin Earth thought-experiment to see why externalism does *not* threaten our *basic self-knowledge*.

[1] Burge, "Individualism and Self-Knowledge."

Suppose that our thoughts about the environment are what they are due to the nature of the entities to which our thoughts are causally linked. According to the standard Twin Earth thought-experiment, a person in the same mental state on Earth and Twin Earth would have different thoughts if the two environments were different in some relevant respect. One could not tell by mere introspection if he was having Earth-thoughts or Twin Earth-thoughts. If the individual were unknowingly switched between the 'home' and 'foreign' situation he could not tell which environment his thoughts were about, *and* he would not *feel* any different. Furthermore, his thoughts *would not* switch: his thoughts about 'water' on both Earth and Twin Earth would be about the substance $water_E$ (where Earth is the 'home' situation). If he stayed in each environment long enough to adopt the relevant concepts (slow switching), his thoughts *would be* different: on Twin Earth his thoughts about 'water' would be about $water_{TE}$ and on Earth they would be about $water_E$.

This does not, however, suggest that the individual would be able to tell the difference between the two environments or that he could tell the difference between his $water_E$-thoughts and his $water_{TE}$-thoughts. It does suggest that there is something about the nature of one's thoughts such that some aspect is fixed by the chemical composition of the individual's body. Burge calls these aspects *pure phenomenological feels*. In other words, there is a completely internal aspect to one's mental events that is impervious to external environmental factors.

It is absurd, says Burge, to move from the fact that an individual who has undergone a series of slow switches cannot tell the difference between his 'home' and 'foreign' thoughts, to the conclusion that he could not know the content of his thoughts without performing an empirical investigation of his environment. According to the reliabilist theory of epistemic justification, one can individuate one's thoughts without subjecting the individuated thought to

some set of externally determined criteria. The conditions that exist for one to have a first-order thought that something is water must obtain if one is to claim that 'this *x* is water.' That is, external conditions are necessary for the individual to bear a proper causal relation to water if he is to make such a knowledge claim. But *knowledge* of this causal relation is not necessary for one to *think* about or have thoughts about water. The complex conditions that must hold for water to be in the environment need not be empirically investigated; they simply must obtain.

So, what does it mean to have knowledge of one's thoughts if the conditions for determining the content of one's thoughts need only be presupposed? The short of it is that knowing what one is thinking requires only the ability to think. Basic self-knowledge judgments are second-order in nature. They depend on the first-order thought for their content but in such a way that thinking *that p* is not merely an object of thought, it is "thought and thought about in the same mental act." (Burge, 472) The 'enabling conditions' of *p* must simply be satisfied; they need not be known.

Burge discusses a parallel between perceptual knowledge and self-knowledge in an effort to clarify why it is that the $same_L$ relation, if you will, need not be known for the individual to know the content of his thoughts. He reminds us that when it comes to perceptual knowledge, we do not require the individual to check for all possible counterfactual situations to ensure that the enabling conditions that verify perceptual claims do obtain. When one claims that he sees food on the table, for example, we do not require him to check the light source for possible mirror-induced optical illusions; nor do we require that he check for counterfeit food in the area that might increase the odds that the food he reported seeing is not real food. Indeed, the objectivity of perception is grounded on the assumption that a perceiver's beliefs, dispositions, and perceptions are not infallible. The "very nature of objective perception insures that the

perceiver need not have a perfect, prior mastery over the conditions for his perceptual success." (Burge, 473)

Similarly, the reflexive nature of basic self-knowledge grants us our epistemic privilege. We need not determine, through empirical investigation, the conditions that make second-order thoughts possible to know our thoughts. We need only think them in the appropriate self-ascriptive manner. Thus, to know one's own thoughts one does not put the content of the thought through some set of criteria for identifying the thought. Furthermore, one need not compare the thought one is having with another thought one is *not* having in order to individuate it. By simply thinking a thought self-ascriptively one individuates the thought from all others. For example, 'I judge that this liquid is water' is not the same mental event as the thought 'this liquid is water.' The former is a case of basic self-knowledge and does not require empirical investigation; the latter is a perceptual knowledge claim, its truth does hinge on external conditions, and only *its* truth requires empirical investigation.

So far Burge has shown that perceptual knowledge and self-knowledge are analogous with respect to individuation. What is now needed is an explanation of the special epistemic status of self-knowledge in contrast with the non-special epistemic status of perceptual knowledge. This difference in epistemic status is dependent upon the fact that perceptual knowledge is objective, whereas self-knowledge is, in some sense, subjective. Let me spell this out more clearly. There are two notions of objectivity fundamental to perceptual knowledge: in one sense, there is no necessary relation between one's perceptions and the objects of one's perceptions, and in another sense, perceptual knowledge is impersonal.

The first sense of the objectivity of perceptual knowledge suggests that it is always the case that one's perceptions of any particular object could be mistaken since the perception and

the object of perception are 'fundamentally independent.' Furthermore, this independency allows the possibility that the nature of the physical entity is different even while one's perceptual states are the same. This entails the fact that we are susceptible to 'brute' errors like misperception or hallucination. That is, we could be mistaken about the objects we perceive. Making a brute error is not indicative of a failure of one's perceptive capacities. Rather, the possibility of such errors brings to light the independence of object and perception. Our perception of objects is not what 'fixes' the object.

The impersonal sense of the objectivity of perceptions is simply an extension of the first aspect. Given any particular circumstance, all individuals have equal epistemic right to make the same observation or perception. In other words, had individual *B* been at the same place at the exact same time as individual *A*, he would have made the exact same observation that *A* actually did make. Equally, the counter-factual observation of individual *B* would have the same justificatory status as the actual observation of individual *A*. There is nothing inherent in the fact that it was individual *A*'s perceptual observation, as opposed to individual *B*'s perceptual observation, that made it justified.

Basic self-knowledge is different from perceptual knowledge with respect to both of these senses of objectivity. First, the object, or subject matter, of one's first-order thought is contingent upon the external world; second, the content of the second-order judgment one has about a first-order thought is dependent upon the first-order thought. A 'gap' between one's second-order thoughts and the subject matter is simply not possible because these thoughts are self-referential and self-verifying. Any error in such cases indicates something wrong with the thinker.[2]

[2] See 3.2 for a criticism of this claim and 3.3 for a defense of it.

Moreover, cases of basic self-knowledge are 'essentially personal' in that their special epistemic status as directly authoritative fundamentally depends on their being made from and about one's first-person point of view. For instance, when I make any judgment, the point of view and the time of the judgment must be the same as the thought I am making a judgment about. The first-person pronoun used in the basic self-knowledge claim indicates that the point of view of the judge and the thought being judged are the same. Burge contends that these differences ground his claim that it is even less plausible in the case of self-knowledge than in the case of perceptual knowledge that knowledge of the conditions that make self-knowledge possible is required.

Burge asks us to imagine a case of slow switching between 'actual home,' where the person thinks that $water_E$ is a liquid, and 'actual twin-home,' where he thinks $water_{TE}$ is a liquid. The individual has acquired the concepts relative to each 'home' though he cannot determine when to use the correct reference. Burge tells us that the person is right and fully justified in both cases even though $water_E$ is found only on 'actual home' and $water_{TE}$ is found only on 'actual twin-home.' Moreover, if the person was told that the switches had occurred and asked himself, "Am I now thinking about $water_E$ or $water_{TE}$?" he would have to answer, "Both," because both concepts are being used. In either of these situations, given "that the thought is fixed and that the person is thinking it self-consciously, no new knowledge about the thought could undermine the self-ascription – or therefore its justification or authority." (Burge, 476) In other words, the content of the first-order thought is fixed non-individualistically, but the second-order thought, by its reflexive, self-referential nature, contains and takes as its subject matter the content of the first-order thought. Counterfeits are logically impossible for self-referential, second-order thoughts, assuming one has not made a 'brute error' or that one is not suffering

from some psychological deficiency; one does not, therefore, need to master the enabling conditions of his knowledge claims.

Burge warns us that we ought not think of self-knowledge as a kind of 'perfected' perceptual knowledge because objects of self-knowledge, i.e. first-order thought contents, need not be fully understood in order to know them. That is, given reliabilism, we need not know all the enabling external conditions that make first-order content possible. Burge tells us:

> The source of our strong epistemic right, our justification, in our basic self-knowledge is not that we know a lot about each thought we know we have. It is not that we can explicate its nature and its enabling conditions. It is that we are in the position of thinking those thoughts in the second-order, self-verifying way. Justification lies not in the having of supplemental background knowledge, but in the character and function of the self-evaluating judgments. (Burge, 477)

In other words, thinking of the content of basic self-knowledge claims as something like the physical objects of perception suggests that there is some level of information that must be met for the claim to be justified. Self-knowledge is different from perceptual knowledge in exactly this respect: self-referential, second-order thoughts have a subjective nature, whereas perceptual knowledge or first-order thoughts have an objective nature. This point is made more explicit in Burge's discussion of the personal nature of self-knowledge as opposed to the impersonal nature of perceptual knowledge.

One reason we might be tempted to think anti-individualism is incompatible with authoritative self-knowledge is that we often waver between the first-person point of view we take when having a thought and the third-person point of view we take when evaluating our thoughts. That is, we take the first-person point of view when we have a thought, but we can imagine, from an omniscient third-person point of view, instances when our thought is not true. For example, we may have thoughts about $water_E$, but taking the omniscient third-person point of view we find out that we are on Twin Earth where there is no $water_E$ to think about. From this

third-person perspective we come to doubt our original first-person authority with respect to our own thoughts. "[We] are easily but *illegitimately* seduced into the worry that our original first-person judgment is poorly justified unless it can somehow encompass the third-person perspective, or unless the third-person perspective on empirical matters is irrelevant to that character of the first-person judgment." (Burge, 478 my emphasis)

Since the justification of Burge's cases of *basic self-knowledge* requires only that the individual self-ascribe a self-verifying second-order judgment, there is no reason to require one to explicate one's thoughts correctly in order to know that he is having those thoughts. It is not the case that a person has first-person authority over how his thoughts are to be explicated or individuated. As we have seen, we do not require conceptual explication for the justification of perceptual knowledge claims. The first-person authority, or the privileged epistemic status of the second-order judgment is simply that it is self-verifying; it does not require further empirical investigation to know that it has been thought.

In brief summary, then, the Burgean paradigmatic cases of self-knowledge, *basic self-knowledge*, bring to light the compatibility of the epistemic privilege we have with respect to our own thoughts and the externalist thesis that claims the content of our thoughts depends, in part, on our relation to the external world. Given epistemic reliabilism, self-knowledge is similar to perceptual knowledge in that complete empirical knowledge of the external conditions necessary for the existence of thoughts and perceptions is not necessary for our ability to have those thoughts and perceptions. The privileged authority of basic self-knowledge is the result of the second-order judgment being formed in a self-referential, self-verifying manner.

3.2 BOGHOSSIAN'S CRITICISM OF BASIC SELF-KNOWLEDGE

Paul Boghossian offers two powerful objections to Burge's *basic self-knowledge* in his essay "Content and Self-Knowledge." First, he claims that Burge's paradigm of self-knowledge cannot account for the general phenomenon of self-knowledge; second, self-knowledge claims in Burge's *basic* sense are not genuine cases of knowledge. If Boghossian's claims are correct, Burge's compatibilist strategy for reconciling self-knowledge and externalism is in serious trouble. Before tackling Boghossian's charges let us recapitulate the specific requirements Burge places on cases of basic self-knowledge.

Burge tells us that only those self-referential and self-verifying judgments are cases of basic self-knowledge. These judgments are self-verifying only if they are exactly coincidental with the thoughts being judged. Thus: "I judge: I think writing requires concentration." The judgment is known non-inferentially because it is about and coincidental with the thought "I think writing requires concentration." No investigation of the environment is needed to know the second-order judgment about the first-order thought. In other words, the second-order thought is known directly and authoritatively fulfilling the very basic requirements for self-knowledge claims. Burge believes that this specific, paradigmatic case of self-knowledge is enough to thwart the threat the externalists levy against the possibility of knowing our own thoughts in a direct (non-inferential) and authoritative manner.

Boghossian takes issue with this last Burgean claim. The general phenomenon of self-knowledge, says Boghossian, includes many more cases than those proposed by Burge. That is, we generally think we know our *standing* beliefs and desires in a direct and authoritative manner, but Burge's proposal does not account for this possibility. Take, for instance, the following two judgments, which are prohibited from being cases of basic self-knowledge:

I judge: I *believe* that writing requires concentration,
and,
I judge: I *desire* that writing require concentration.

According to Boghossian, I need not believe or desire that writing require concentration in order to *judge* that I believe or desire that writing require concentration. "This would appear to be a serious problem [since] we do know about our beliefs and desires in a direct and authoritative manner, and Burge's proposal seems not to have the resources to explain how." (Boghossian, 495)[3] In other words, if I can judge that I believe or desire that *p* without actually believing or desiring that *p*, then the judgment is not self-verifying. The act of making the judgment itself does *not* make the judgment true contrary to Burge's claim that making the judgment itself *does* make it true.

Similarly, Burge's basic self-knowledge fails to account for *occurrent* mental events such as: I judge: I fear that writing requires concentration. Again, I need not actually fear that writing requires concentration in order to make the judgment that I do so fear. The judgment regarding my fear is not self-verifying since I do not fear that writing requires concentration. Perhaps the best possible cases for being basic self-knowledge are the mental events with which Burge restricts himself, i.e. judgments about and coincidental with thoughts. But even these fail to account for the *general* phenomenon of self-knowledge, according to Boghossian, since we normally believe we know thoughts that we have had in the past. Such knowledge claims are "central to our capacity for self-knowledge." (Boghossian, 496) That is, Burge's basic self-knowledge severely limits the types of thoughts we can know directly and authoritatively. He suggests that only those self-verifying judgments can be known non-inferentially; and for a judgment to be self-verifying it must occur coincidentally with the first-order thought it judges.

[3] Boghossian, "Content and Self-Knowledge."

However, our common experience is that we often do know, directly and authoritatively, the thoughts we have had in the past, however minimally we construe the past. That is, suppose you have a thought, *X*, at time t_1. According to Burge, the only way you could non-inferentially know that thought is if it were accompanied by a judgment about it, which also occurs at t_1. Common experience shows that we often do know, after the fact, those thoughts we had at t_1, without those thoughts being accompanied by coincidental judgments. Suppose at t_1 you think, "I will eat a hamburger for lunch today." Then at some later time, say t_2, you judge: "I had the thought that I will eat a hamburger for lunch today." It seems that you do know, at t_2, the thought you had at t_1, despite the fact that the thought you had at t_1 was not coincidentally accompanied by a second-order judgment about it. The second-order judgment was about the first-order thought, but it did not occur until t_2. "The fact that, *had* the thought been part of a second-order judgment, then that judgment would have been self-verifying, does not help explain how we are able to know what thought it was, given that it *wasn't* part of such a judgment." (Boghossian, 496)

For Boghossian's second objection that Burge's basic self-knowledge are not instances of genuine knowledge, let us return to the Twin Earth thought-experiment in which a subject, *S*, undergoes a series of slow switches between Earth and Twin Earth. Recall that experiencing the switches 'slowly' allows *S* to acquire the concepts appropriately relevant to each environment, though he is unaware of using different concepts.[4] Burge claims *S* is as much right and justified in his Earthian thought, "I am thinking that water_{E} is a liquid," as he is in his Twin Earthian thought, "I am thinking that water_{TE} is a liquid," precisely because knowing that a switch has

[4] Please note that Boghossian is granting the possibility of the existence of Twin Earth simply in an effort to show that the Twin Earthian concepts are relevant alternatives for *S*, despite *S*'s ignorance of them, since a requirement for knowledge is that the subject rule out all relevant possibilities. Hence, knowing that the liquid in the glass is water and not gin requires that both water and gin be relevant alternatives in S's environment.

occurred is "irrelevant to the truth and justified character of these judgments." (Burge, 476) The true and justified character of such judgments is a result of their being self-referential and self-verifying. Now, if *S* were told of the switching he would not know non-inferentially whether he was thinking about $water_E$ or $water_{TE}$ yesterday, since that knowledge would require knowing which environment he was in yesterday and may require a "complex story" about memory.

Boghossian finds these Burgean claims rather mysterious and unsatisfying. First, Burge appears to be saying that *S* is in a position, at t_1, to have a self-verifying judgment about his t_1-thought, which grants *S* direct and authoritative knowledge of his thoughts. Then, Burge claims that *S* is not in a position, at t_2, to have a self-verifying judgment about his t_1-hought. To know his t_1-thought at t_2, *S* must investigate his environment, which does not allow him to have direct and authoritative knowledge of the t_1-thought. This conclusion seems to fly in the face of our ordinary conception of memory: "if *S* knows that *p* at t_1, and if at (some later time) t_2, *S* remembers everything *S* knew at t_1, then *S* knows that *p* at t_2." (Boghossian, 497) According to Burge, however, it is not the case that *S* knows that *p* at t_2. There are only two explanations: either *S* forgot, at t_2, that *p*, or *S* never knew that *p*.

It does not seem to be the case that externally individuated thoughts are easy to know but hard to remember. There does not appear to be anything about the nature of 'relationally individuated content' that would suggest such content is more susceptible to memory failure than directly knowable thought contents. So, we can rule out the first option 'by stipulation,' as Boghossian puts it. We are then left with the conclusion that *S* never knew that *p*. If this is true, then Burge's basic self-knowledge, his self-verifying judgments are not instances of genuine *knowledge* and so cannot be used to disarm the externalist threat to self-knowledge.

I would now like to offer a possible criticism of Boghossian's first objection to Burge's basic self-knowledge. It is not entirely clear how instances in which one judges that he believes or desires *that p*, but does not actually believe or desire *that p*, are central to our capacity for self-knowledge, that such claims are representative of the general phenomenon of self-knowledge. It seems, rather, that they are clearly cases in which the individual does not know what he thinks he knows. For example, suppose Jones claims, "I judge: I believe I am sixty years old," though, in fact, Jones is only forty years old. Clearly, either Jones has a false belief, or he does not actually believe that he is sixty years old. I am willing to grant that it is possible that Jones has a false belief about his age. If he does not actually believe the content of his second-order judgment, then why should we think his judgment is a knowledge claim at all? Surely, one who makes such mistaken judgments, in either case, is not operating in a proper fashion. That is to say, such an individual would not seem to be in any normal mental state, at least not in the sense of an appropriately functioning mental state. Jones might be suffering from some form of self-deception or be psychologically deficient in some other respect. We would not normally say that such an individual *knows* what it is that he thinks, regardless of whether the thought was accompanied by a second-order judgment.

If I were to say, "I judge: I believe that writing requires concentration," but in fact I do not believe that writing requires concentration, my judgment is not self-verifying, as Boghossian has explained. But this fact also shows that my judgment is not a knowledge claim at all. Therefore, it does not seem that this is a more paradigmatic case of self-knowledge than Burge's self-verifying judgments. It may seem that I am suggesting that self-knowledge must be infallible if it is to be knowledge. However, I am perfectly willing to grant that a judgment like, "Jones judges: I believe that water is XYZ," is self-verifying, not because water is XYZ, but

because Jones *does* believe that it is. My belief may be mistaken, but my judgment about my belief is not. The fundamental difference between the previous judgment and those Boghossian considers is that the individual must actually hold the belief or desire if it is a knowledge claim. I think there is an implicit requirement in Burge's construal of basic self-knowledge that the thought actually be thought. If we apply this implicit requirement to all *standing* and *occurrent* mental states, I think Burge's basic self-knowledge can account for the general phenomenon of self-knowledge.[5]

Boghossian's second objection to Burge at first seems more plausible, and so, more devastating. I agree that the requirement that the second-order judgment be coincidental with the first-order thought is too restrictive, since it does seem that knowledge of past thoughts can be known to the individual directly and authoritatively, though not infallibly. I do not, therefore, agree with Burge that the self-verifying nature of self-knowledge is contingent only upon a second-order, coincidental judgment about a particular thought, belief, or desire. Nor, however, do I agree with Boghossian's resignation that we are stuck in a quandary. Forget Twin Earth for a moment and think more concretely with me.

Suppose I think, at t_1, "My desk is brown." At some later time, t_2, I say, "At t_1 I thought that my desk is brown." The truth of the color of my desk is externally determined. More importantly, however, is that what is at stake is whether the fact that I was thinking of my desk is externally determined. According to the ordinary conception of memory, so long as I remember everything I thought at t_1 I will know at t_2 what I thought at t_1. With Boghossian, I believe that even if the content of my thought about my desk is externally determined, there is nothing about the nature of the external determinant of my thought about my desk that makes it particularly difficult for me to remember. So, regardless of whether I am, at t_2, in the presence of my desk, I

[5] For a criticism of this claim see Sec. 3.3.

should have no trouble remembering that it was my desk that I had a thought about. There is no reason to suspect that I have forgotten that I thought that my desk was brown. Counterfeit desks may be present in my vicinity at t_1 or at t_2, but the thought I had about my desk is as much directly and authoritatively knowable to me at t_2 than it was at t_1.

Now, let us return to the Twin Earth thought-experiment, which suggests that unless I knew that it was *my* desk, and not a twin desk, that I thought was brown, I could not know my thought directly and authoritatively. Knowing which desk I was referring to, according to the thought-experiment, would require knowledge of the environment in which I had the thought that my desk is brown. Though the content of my thought may be externally determined, the fact that I had a thought is not. Knowing that I had a thought is not externally determined precisely because the environment in which I had the thought did not determine *that* I thought. The environment does verify the truth of the content of my thought, but it cannot verify that I had a thought in the first place. My thought at t_2 is not *coincidental with* my thought at t_1, and so is not self-verifying. But, if my t_2–thought is indeed *about* my thought at t_1, it is an instance of self-knowledge because, with Burge, I contend that I need not know the external conditions that must obtain for me to have a thought in order to know that I indeed did have a thought.

3.3 BERNECKER'S OBJECTION TO BURGEAN COMPATIBILISM

In "Externalism and the Attitudinal Component," Sven Bernecker poses a new, two-part incompatibilist challenge to Burge's compatibilist strategy for reconciling externalism with self-knowledge. He suggests, first, that though Burge's compatibilism 'convincingly' shows that it is *that p* that I believe, it does not show how I can have privileged knowledge that I *believe*, rather than, say, *suppose, doubt, expect*, etc, *that p*. Bernecker claims, "Mental states represent the

union of an attitude and a content." (Bernecker, 502)[6] Full self-knowledge of one's mental states, then, requires knowledge of both these components. Second, given externalism, knowledge of one's attitude is susceptible to a kind of empirically discoverable error and is therefore not privileged like knowledge of the current content of one's thoughts. In other words, Bernecker suggests that when one knows the attitude one takes with respect to one's thought content, one's knowledge of this attitude cannot be the result of mere introspection. Thus, complete privileged self-knowledge is incompatible with externalism.

Let me explain how Bernecker arrives at this incompatibilist conclusion. He begins with a brief summary of Burgean compatibilism. First, according to externalism, in order to *know that p* I must know that the external conditions that make *p* possible actually obtain. However, in order to *think that p* I need not know such information. For Burge's basic self-knowledge, I need only judge that I think *that p* in order to possess privileged self-knowledge of *p*, since the content, that I think *p*, is contained in the second-order, self-verifying judgment. Second, the external conditions for *knowing that p* are the same as the conditions for *thinking* that *p*, furthermore, these conditions are also required for knowing (or believing) that one is thinking *that p*. Third, and finally, according to Burgean compatibilism, with respect to both *knowing* and *thinking that p*, the external conditions must obtain, though only in the case of knowing *that p* is one required to know that the conditions obtain. This last component of Burgean compatibilism results from epistemic reliablism, which claims that a belief must be produced by a reliable process regardless of whether the individual is aware of the conditions that must obtain for the process to be reliable. So, "self-knowledge does not require investigation of one's environment, because the content of the first-order thought is automatically contained in the second-order thought, and the contents of both thoughts are determined by the same causal relations of which

[6] Bernecker, "Externalism and the Attitudinal Component."

one may be ignorant." (Bernecker, 501) Since the content of the second-order judgment is contained (or included) in the first-order thought, Bernecker refers to Burgean compatibilism as the 'inclusion theory,' which seems to presuppose privileged knowledge of the attitude expressed in the first-order thought.

The particular line of criticism Bernecker pursues hinges on the fact that knowledge of one's mental states requires that one know the certain content, *C*, *and* the certain attitude, *A*, of that mental state. One can obviously take a particular attitude, *A*, without having a particular concept of attitude *A*, but one cannot *know* that one has taken attitude *A*, unless one possesses the concept of *A*. Bernecker is quick to tell us, along reliabilist lines, that possessing the concept *A* is not the same as being able to explicate all the necessary application conditions involving *A*. However, "the possession of attitude concepts involves some kind of 'cognitive achievement'." (Bernecker, 502) Presumably, being able to distinguish what appears to me to be the way things are from the way things actually are, is such a 'cognitive achievement,' and is so in some privileged sense.

Bernecker proposes a thought-experiment designed to show the logical independence of content-identification and attitude-identification. Suppose there is a thermometer placed in a gas tank. This thermometer provides (first-order) content information about the temperature of the gas, but it does not possess any attitude toward that information. Now, suppose we have attached a sensor to the thermometer that triggers an alarm when the temperature of the gas is 30° C. The state of the alarm (either on or off) contains content about the thermometer's state, which represents the temperature of the gas. This is a case of what Bernecker calls second-order content inclusion, or in Burgean terms, a second-order self-verifying 'judgment.' Note the parallel to Burge's compatibilism in which I have a first-order thought and a second-order

judgment the content of which is the content of the first-order thought. Both the content of my judgment and the content represented by the alarm's state are reflexive. However, in the thermometer/sensor case, the representational system is unable to have propositional attitudes toward the content it represents; therefore, it cannot possess self-knowledge. In the case of my second-order judgment I do take an attitude toward the content, but Burgean compatibilism does not provide an account of my privileged access to that attitude, so it too fails to provide a full account of privileged access to self-knowledge.

Bernecker supposes that Burgean compatibilism could accommodate this new requirement (that knowledge of the attitude is necessary for complete self-knowledge), but only with respect to *cogito*-like thoughts, or self-verifying second-order judgments. Proponents of Burgean compatibilism wish to extend the inclusion theory to standing mental states like belief, desire, etc. Throughout the remainder of the essay Bernecker demonstrates that extending the inclusion theory to beliefs provides an account of self-knowledge *sans* the privilege compatibilists seek. That is to say, knowledge of one's attitudes is vulnerable to error that can only be discovered empirically. Thus, one cannot have privileged access to complete self-knowledge if we maintain even a minimalist version of externalism, i.e. epistemic reliabilism.

Contrary to the claims I made in 3.2 regarding the impossibility of an individual being able to misconstrue one's attitudes and yet still make genuine knowledge claims (see p. 42), Bernecker contends that the independence of content-identification and attitude-identification grants this possibility. In other words, "one can misrepresent one's attitudes, [if] one possesses incomplete understanding of concepts used to describe one's mental condition." (Bernecker, 505) Bernecker adopts a Burgean thought-experiment to flesh out this possibility.

Suppose Bert thinks arthritis affects both the thighs and the joints, though the medical experts in his linguistic community define 'arthritis' as a disease that only affects the joints. Burge's semantic externalism insists that Bert's usage of the term 'arthritis' refers to the medical expert's definition of arthritis despite the fact that Bert incorrectly uses the term. "Society's use of a term partly determines the concepts of individuals in the society, even of such medically ignorant individuals as Bert." (Bernecker, 505)

Now, suppose Bert's cousin, Oscar, is similarly confused about the concept 'to believe,' such that he has a 'course-grained concept of belief' – his concept of belief is so wide that he takes 'to believe,' 'to suppose,' 'to decide,' etc. to be synonymous, even though his linguistic community defines each of these concepts as not being synonyms. If we adopt Burge's social externalism, we are committed to saying that Oscar's use of his concept 'to believe' is the same as the other speakers in his linguistic community. Thus, if Oscar were to claim that he *believes* arthritis is painful, but in fact he actually *supposes* that arthritis is painful, his mistake is not introspectively knowable. That is, he cannot know that he *believes* that arthritis is painful unless he knows "that the mental state he is in has the kind of features that are constitutive of 'belief,' as his fellow language users employ the term." (Bernecker, 506) *That* knowledge, however, is unavailable to Oscar in any privileged sense: to know that his mental state is constitutive of a belief-state he would have to empirically investigate his social (linguistic) environment. Thus, Oscar does not authoritatively know his attitude and so does not possess privileged self-knowledge about his mental state.

Bernecker suggests two possible reasons for this last claim, rejecting the first and accepting the second. First, one might suggest that Oscar does not possess privileged self-knowledge of his mental state "because his mistaken self-attribution is unremediable by

introspection." (Bernecker, 507) Since Oscar cannot discover that his *belief* is actually a *supposition* simply by 'monitoring his mental condition,' then he is not authoritative with respect to introspectively knowing his attitudes. This line of reasoning, however, does not follow from Burgean compatibilism. In other words, the external component of Burgean compatibilism is epistemic reliabilism, which claims that to know *that p* one does not have to know that one knows *that p*. So on this externalist view, introspective self-knowledge is vulnerable to empirically discoverable error. That is, the fact that Oscar could not introspectively know that his belief was actually a supposition is not what makes his self-knowledge *un*-privileged.

Oscar lacks privileged self-knowledge of his attitudes because he cannot distinguish one attitude from another. His definition of 'to believe' is so broad that "for all he knows he could suppose, decide, or consider" that arthritis is painful. (Bernecker, 507) If Oscar actually does believe that arthritis is painful, his claim is not a knowledge claim because it did not result from a reliable process. Recall that reliablism claims that a belief *that p* must be the result of a reliable process, regardless of whether one knows what that reliable process is, or that he knows that he knows *that p*. Oscar's claim is only accidentally true. Thus, his belief claim is neither privileged nor is it a genuine instance of self-*knowledge*.

Bernecker concedes that since Oscar is a special case and that, in general, we do not suffer from the type of confusion that denies Oscar privileged knowledge of his attitude, the Burgean compatibilists can assert that '*in some sense*' knowledge of our attitudinal components can be privileged. Oddly enough, Bernecker's next claim is that whatever "the account of epistemic specialness of introspective knowledge of attitude might be, it has to differ from Burgean compatibilism." (Bernecker, 508) I must admit that these two claims seem absolutely inconsistent. If the 'epistemic specialness,' or 'privilege,' of our knowledge of our attitudinal

component is not threatened by the Oscar-like cases, given their own 'specialness' or rarity, why must we abandon Burgean compatibilism in order to explain that privilege?

3.4 A THIRD CRITICISM OF BASIC SELF-KNOWLEDGE

Let us turn now to a criticism of Burge's basic self-knowledge that has not been mentioned by either Boghossian or Bernecker. It seems that the externalist can easily take issue with self-verifying second-order judgments on the following grounds. If the content of the so-called 'privileged' second-order judgment is parasitic upon the content of the first-order thought, which is externally determined, why isn't the content of the second-order judgment also, that is, parasitically, externally determined? According to Bernecker's criticism of Burgean compatibilism, it might be the case that empirical investigation is needed to know the attitudinal component of one's thought. Taking this line a step further, it might be the case that knowledge of the second-order judgment also requires empirical investigation, at least in some secondary, parasitic manner.

There is no room for privileged access to the content of one's second-order judgment because the content that judgment is about can only be known through empirical investigation. Despite Burge's contention that the ability to have the first-order thought does not require empirical investigation, both he and Bernecker have demonstrated that *knowing* that the thought occurs might require empirical investigation. The second-order judgment is a knowledge claim about having the first-order thought. Therefore, knowledge of the content of the thought must require both that the individual knows that the thought has occurred (or is occurring, as Burge would insist) *and* knowing the content of the fist-order thought. But this knowledge is parasitic upon the external determination of the first-order thought content. Even Burge's second-order self-verifying judgments, then, lack special epistemic status.

CHAPTER 4 DONALD DAVIDSON

4.1 DAVIDSON'S COMPATIBILIST STRATEGY

Donald Davidson contends that the externalist threat to privileged self-knowledge rests on a faulty picture of the mind. His compatibilist strategy, then, is to discard this 'fundamentally flawed' picture of the mind so that not only will we be in a position to disarm the externalist threat, we will also be able to explain first-person authority in such a way as to make clear the interconnectedness of first-person authority, the social character of language, and the external determinants of thought and meaning. This view is represented in two of Davidson's essays, "First Person Authority" and "Knowing One's Own Mind." Though I will refer to each of these essays throughout this chapter, I take Davidson as presenting one, unified compatibilist theory.

Davidson restates the problem compatibilists face along the following lines. Why is it that I ascribe certain mental states, i.e. beliefs, doubts, etc., to others on the basis of certain evidence, whereas I ascribe those same states to myself without such evidence? What explanation can we give of this asymmetry in the epistemic warrant for ascribing certain mental states to others and the warrant for ascribing those same states to ourselves? (Davidson 1984, 107)[1] One 'explanation' is that even though we generally do not base our self-ascriptions on the same kind of evidence as we do of others, we could. This answer, however, simply gives a description of the asymmetry and does not explain first-person authority. Not only do we not appeal to some evidential support (in fact, we rarely do), we *need* not make such appeals. If we were to appeal to evidential support, we would not have the kind of privilege we presume to have with respect to our self-ascriptions, since whatever evidential support we appeal to is equally available to others. Again, then, how might we *explain* this asymmetry, or the authority we have

[1] Davidson, "First Person Authority."

over our own self-ascriptions, particularly if those states so ascribed depend on external factors for being the states they are?

Though external factors contribute to the determination and individuation of the content of our thoughts, Davidson argues that both Putnam and Burge fail to adequately explain *how* they do so. Putnam sets the problem up in such a way that prohibits a possible solution to the problem of why there is an asymmetry between warrant for first-person self-ascriptions of certain mental states and ascriptions of the same mental states to others. That is, Putnam endorses a false dichotomy: complete subjective and internal content determination, of which one can have privileged or authoritative knowledge, what Putnam calls 'methodological solipsism', on the one hand; and "ordinary beliefs, desires, and intentions, as we commonly attribute them on the basis of social and other outward connections, on the other," over which one cannot have first-person authority. (Davidson 1986, 95)[2] I will return presently to Putnam's externalism and why Davidson thinks it prevents a compatibilist solution. For now, let me explain Davidson's criticism of Burge's social externalism.

Burge, too, seems to 'seriously compromise' first-person authority in his explanation of how external factors determine the content of one's thoughts.[3] According to social externalism, a person can believe the content of his thought even on the basis of a partial understanding of the content. That is, it may be true that I believe that I have arthritis even if I misconstrue the definition of arthritis such that I think it applies only to one specific cause of joint inflammation when the correct definition of arthritis allows for various causes of joint inflammation. Now, according to Davidson, it seems to follow that if a person is partially misinformed about the

[2] Davidson, "Knowing One's Own Mind."
[3] This explanation is found in Burge's essay "Individualism and the Mental," which is *not* the essay discussed in Chapter 3. However, the social externalism explained in "Individualism and the Mental" is presupposed in his essay "Individualism and Self-Knowledge," which was discussed in Chapter 3, so the reader ought to follow this discussion easily.

meaning of the words he uses to express his belief, he must also then be partially misinformed about his belief so expressed. Burge contends, however, that both a misinformed speaker and an informed speaker mean the same thing by the following utterance, "Carl has arthritis." (Davidson 1986, 98) Davidson agrees with Burge that the content of one's thought is not exclusively fixed by what goes on in one's head, but he rejects Burge's claim as to how the social dimension of language determines the meaning of a misinformed person's belief.

In other words, according to Davidson, it is a mistake to *always* assume that the intended meaning of one's word is the same as the socially determined meaning of the same word. When I claim that Carl has arthritis, it is possible that I do so on the understanding that he has calcium deposits in his joints. Suppose you know that Carl has the gout in his joints (according to the 'correct' social definition of arthritis, gout is an actual cause of arthritis). My use of the word 'arthritis' is intended to mean one thing; your use of the word 'arthritis' is intended to mean another. Burge contends that 'arthritis' has the same meaning whether uttered by an informed speaker or a misinformed speaker. Davidson thinks it is absurd to make such a claim, since clearly you and I intend something different in our ascription of arthritis to Carl. Now, it is true that Carl has arthritis, and so technically your and my belief that Carl has arthritis is true, but there is a 'relevant difference in the thoughts' you and I expressed, i.e. I believe Carl suffers from calcium deposits in his joints, you believe Carl has the gout. Therefore, says Davidson, we ought to reject Burge's construal of the manner in which social factors determine the meaning of words (and thoughts, beliefs, etc.) without rejecting altogether that such meanings are socially determined. That is, according to Davidson, the natural history of a word, i.e. when it was learned and used, is the only proper explanation of how social factors determine meaning. What is left, then, is whether this manner of social determination threatens first-person authority.

Davidson suggests that the manner in which external, social factors determine the meanings of one's words, and therefore the content of one's thoughts (beliefs, desires, etc.) is dependent upon the requirements of 'interpretability.' This suggestion appears to be in the direct lineage of Wittgenstein's argument against the possibility of a private language. Wittgenstein claimed that any language one employs is in principle knowable to any other language speaker. Thus, one cannot make up a language such that, in principle, no one else could ever know it. Similarly, Davidson argues that someone cannot mean something by his words that someone else could not decipher. Since the purpose of language is to communicate effectively, a speaker must make every effort to make his words intelligible to another speaker. Therefore, the 'irreducible social factor' is interpretability, not, as Burge contends, general social usage. (Davidson 1984, 100)

Let us return, quickly, to Putnam's Twin Earth thought-experiment, which showed that two people could be in exactly similar physical, and therefore mental, states and yet mean two different things by the same word. Putnam contends that this fact suggests that neither individual can know what he thinks or means, first, because of Putnam's assumption that "if a thought is identified by a relation to something outside the head, it isn't wholly in the head;" and second, because "if a thought isn't wholly in the head, it can't be grasped by the mind in the way required by first person authority." (Davidson 1986, 102) Davidson argues that the first assumption is suspect: it does not follow from the fact that meanings are externally individuated that they are not in the head.

In other words, Putnam takes his Twin Earth thought-experiment as evidence against identifying mental states with physical states, since doing so (e.g. the case of $Oscar_1$ and $Oscar_2$) does not yield knowledge of the meaning of one's words, and by extension, nor does it yield

knowledge of one's thoughts. Burge also attacks mental-physical identity theories. He argues that the conclusion that a physical event (say, a neural event) is not identical with an individual's thought does not follow from the intuitive implausibility of the denial of the claim that no thoughts could have different content and yet be the same token event to the conclusion. He claims that materialist identity theories, i.e. mental-physical identity theories, suppose that the content of a mental event (state) can vary while the mental event itself remains constant. This picture of the mind is false, according to Burge (but in Putnam's terminology), because when $Oscar_1$ has a thought about $water_E$ described in terms of physiology, biology, chemistry, etc. as mental event *A*, then the thought $Oscar_2$ has about $water_{TE}$, cannot also be described as the same physical event, *A*, that $Oscar_1$ is in since the extension of their words, and therefore the contents of their thoughts, are different.

According to Davidson, however, the existence of external determining factors is not enough to discredit mental-physical identity theories, despite both Putnam and Burge's contention that it is. Davidson contends, with Burge, that two mental events with different contents must certainly be different events. Putnam and Burge's thought-experiments are meant to demonstrate this fact: two individuals can be exactly similar in all relevant physical respects and yet can differ in what they mean or think. They contend that, since $Oscar_1$ and $Oscar_2$ cannot have the same physical event despite having the 'same' thought, their thought-experiments prove that mental states are not identical with physical states. Davidson claims that this conclusion does not follow simply because "there is *something* different about them, even in the physical world; their causal histories are different." (Davidson 1986, 104) They learned the meaning of the words they use in different social contexts.[4]

[4] This claim is suspect; but I will defer my criticism of it to section 4.3.

Thus, when my Twin Earthian twin utters something about 'water,' one must interpret him differently than when I utter something about 'water.' "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 knew that he holds a sentence true, he knows what he believes." (Davidson 1984, 111) Despite the fact that my twin and I may not know if we are referring to H_2O or XYZ, given this presumption regarding the nature of interpretation, both my twin and I have first person authority over our thoughts about 'water.' That is, we generally know the meaning of the words we use since those words were learned in a social environment. It is true that we may have been misinformed about our words' meanings or that we only partially understood the meanings, but, according to Davidson's picture of interpretation, an interpreter must presume that we have first person authority over our thoughts since we do not generally second-guess the meanings of our words.

Davidson argues that Putnam's conclusion that external determination of meaning and thought content precludes the possibility one's first person authority over one's mental states such as beliefs, desires, etc. is the result of working within a 'fundamentally flawed' framework. Putnam's picture of the mind posits mental states or propositional attitudes as 'objects of thought': actual entities that the mind can 'grasp,' 'entertain,' or 'have before' it. If this were true, it is easy to see how external determination or individuation of those mental states might be problematic for first person authority. "For if [to] be in a state of mind is for the mind to be in some relation like grasping an object, then whatever helps determine what object it is must equally be grasped if the mind is to know what state it is in." (Davidson 1986, 106) 'Grasping' the external determinants of a mental state, then, could not come about from any privilege or special epistemic status of the person in that mental state. Thus, Putnam holds that there are

certain 'inner' states that one can have authority over, but there are also other ordinary states like belief, desire, meaning, etc. that are only knowable through their social determinants.[5]

There is an easy way to free ourselves from the incompatibilist conclusion Putnam would have us concede. We must rid ourselves of the metaphor of mental states or propositional attitudes as objects given to the mind. After all, says Davidson:

> [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 object it is. (Davidson 1986, 108)

Though a person is said to *have* beliefs, doubts, and desires, these beliefs, doubts, and desires are not actual entities; moreover, having these mental states or propositional attitudes does not require one to have a corresponding object that must be grasped by the mind in order for the mind to know it. Putnam, Burge, and others are right, according to Davidson, in concluding that no object could fulfill the dual requirements of being "'before the mind' and also such that it determines what the content of a thought must [be]." (Davidson 1986, 108) So, the only way out of our quandary, the only way to explain the actual asymmetry between an individual's epistemic status with regard to his own mental states and others' third-person epistemic status is to think of first-person authority in terms of the nature of interpretation. We should simply discard the metaphor of 'objects of thought' altogether, which will free us to presume that individuals do have privileged access to, or first-person authority over their mental states, the content of their thoughts, their propositional attitudes, or whatever you want to call them.

However, as Davidson tells us, we must explain how this privilege is compatible with externalism. It is not enough to simply describe the asymmetry between my own warrant for my self-ascriptions and your warrant for ascribing mental states to me. Davidson's explanation, as

[5] It is left unsatisfyingly unclear in Davidson (1984 & 1986) and Putnam (1975) what these 'certain inner states' are.

we have seen, is given in terms of the nature of language. Languages are tools for effective communication; their most important requirement is interpretability. Using words whose meaning one has learned through social factors enables others to understand what one is saying. Further, if two people did not share a common language, the speaker must speak with consistency if the other, non-speaking interpreter is to be able to understand the meaning of the speaker's words. In either case, the interpreter must presume that the speaker knows what he means by the words he employs. The speaker, as well, need not question whether his words mean what he thinks they mean, simply because "whatever [he] regularly does apply them to gives [his] words the meaning they have and [his] thoughts the contents they have." (Davidson 1986, 109) The social determination of the meanings of words comes into play early in the speaker's linguistic life. Once one learns a language, however appropriately or accurately, one is free to use that language to refer to objects in the world or to express one's own thoughts.

Of course, the speaker may be wrong about the meaning of his words. First-person authority is not to be construed as completely authoritarian precisely because error *is* possible. Virtually no one, compatibilist or incompatibilist, contends that indubitability is an essential feature of first-person authority. Though the speaker is susceptible to error, he is not likely to misconstrue what *he* means by his words.[6] Indeed, no matter what the speaker means, there will always be an external determinant of the truth of a statement. "The speaker, after bending whatever knowledge and craft he can to the task of saying what his words mean, cannot improve on the following sort of statement, 'My utterance of "Wagner died happy" is true if and only if Wagner died happy.'" (Davidson 1984, 110-11) Since the speaker need not interpret his own words, his warrant for knowing what he means or believes will always be different, in the sense

[6] See the end of section 4.3 for a criticism of this claim.

of being privileged, from the warrant anyone else has for ascribing meaning to his words or the contents of his beliefs.

4.2 AN OBJECTION TO DAVIDSON'S COMPATIBILISM

One might argue that Davidson has not provided an adequate compatibilist account: inasmuch as he has endorsed a mental-physical identity theory, there is no room for privileged access or first person authority.[7] That is, from the notion that mental states are physical states it seems reasonable to count mental or psychological states as neural events. Neural events are physical events and all physical events are publicly observable. Therefore, psychological states must be publicly observable. Given first-person authority, which claims that individuals have privileged access to their own psychological states, it appears that Davidson cannot endorse both a mental-physical identity theory and first-person authority. Psychological states cannot both be privileged and publicly observable. Let me illustrate this objection with an example.

Suppose Jones is in psychological state *S*. According to the objection, if Jones' knowledge that he is in state *S* is privileged, it must be the case that both only Jones could know it and Jones knows he is in state *S* simply on the basis of introspection. Now, suppose that all psychological states are neural events, which can be mapped and displayed on a computer screen thanks to recent technological advancements in brain studies. Obviously, given this technology, psychological states are publicly observable, since anyone with the visual capacity to see the computer screen will be able to have knowledge of Jones' psychological state. Jones'

[7] Sarkar, "Three Counter-Arguments: A Rough Sketch," p. 4, states the objection as follows:
[1] Psychological states are neural events. (Davidson 1986, 104)
[2] Neural events are physical events.
[3] Physical events are publicly observable.
[4] Psychological states are publicly observable. (From [1], [2], and [3])
[5] Individuals have privileged access to their psychological states, i.e. psychological states of an individual are not publicly observable. [Thesis of *first person authority*. Isn't this what Tyler Burge means when he refers to an experience as being person? Pp. 477-478]
[6] Therefore, either [4] or [5], but not both.

knowledge of his psychological state, therefore, is no more privileged than anyone else's knowledge of Jones' psychological state. This conclusion follows directly from Davidson's contention that mental-physical identity theories are admissible. Therefore, Davidson has not presented an effective compatibilist strategy precisely because first person authority cannot find footing if a mental-physical identity theory is in play, or so the objection might run.

4.3 AN ATTEMPTED DEFENSE OF DAVIDSON'S COMPATIBILISM[8]

I think the previous objection to Davidson's compatibilist strategy suffers from two serious problems. First, the objection only holds given a strong reading of Davidson construal of mental-physical identity theories; and second, it misconstrues what 'privileged self-knowledge' or 'first person authority' entails. If either of these problems can be demonstrated, Davidson's position might seem more plausible; at the very least, it should seem more consistent than the objection gives it credit.

The first problem with the objection is that a weak reading of the following claim shows that Davidson might *not* be suggesting that psychological states are physical (neural) events. He claims: "I see no good reason for calling all identity theories 'materialist'; if some mental events are physical events, this makes them no more physical than mental. Identity is a symmetrical relation." (Davidson 1986, 104) Mental events are 'identical' with physical events in the weak sense of being represented by physical events. 'Brain mapping' may represent the mental state Jones is in, but this does not suggest that the mapping on the screen *is* the mental state.[9]

Consider an analogy. Modern stereos are equipped with equalizers that allow the listener to set the various levels of bass, treble, and mid-range frequencies. The levels are digitally

[8] A complete defense of Davidson must take into account Davidson's views on 'anomalous and neutral monism.'
[9] The strong reading of the claim that "if some mental events are physical, this makes them no more physical than mental" is that the physical and mental aspects of a thought are like two sides of the same coin. This reading allows the objection to go through because it denies the idea that the physical aspect merely *represents* the mental aspect.

represented on an LCD display mounted on the front of the stereo. When music is played through the stereo, the equalizer digitally represents the various frequencies of the recorded music. Surely we would not say that the digital representations *are* the frequencies. What I take Davidson's claim to suggest is that the physical representation of a psychological state is no more the state itself than the digital representations of frequencies on an LCD display are the frequencies themselves. Therefore, the objector is wrong to suppose Davidson is committed to the idea that psychological states *are* neural states. Davidson's compatibilism should survive if we commit him only to the weaker claim that psychological states are represented by neural events. Now, the objector might insist that even this weaker claim undercuts Jones' first person authority, which brings us to the second problem the objection faces.

First-person authority, or privileged access, describes the manner in which the individual in certain mental states comes to know that he is in those mental states. Davidson's restatement of the problem compatibilists face is explicit about this point. His compatibilist strategy is an effort to explain the asymmetry between the way in which an individual self-ascribes certain mental states and the way in which others ascribe those *same* mental states to him. Nothing Davidson (or Burge) has said about first-person authority or privileged access suggests that the self-knowledge claim is, in principle, knowable only to the individual in the state in question.

Rather, the manner in which an individual in state *S*, say, Jones, knows that he is in state *S* is different from the manner in which anyone else knows that Jones is in state *S*. Regardless of whether a psychological state is a neural event in the strong sense of actually being a physical state, or, to put it more weakly, if a psychological state is simply represented by a neural event, the manner in which Jones knows that he is in state *S* is different from the manner in which you

or I know that he is in state *S*. His knowledge is privileged because he need not, in principle, look at the computer-generated representation of state *S* to know that he is in it.

The objector might insist that there is a sense in which knowledge of my thought is privileged such that no one else can, in principle, know the thought I have or even know that I had a thought. Suppose I have a thought but do not express that thought in any way. Certainly, I am the only one who could possibly know (1) what thought I just had, and (2) that I had a thought at all. If we presume that this is the only instance of privileged self-knowledge or first-person authority, then claiming mental states are neural (physical) events will undoubtedly undermine my first-person authority. In principle, if my thoughts are neural events, they can be publicly observable, and so my knowledge of my thoughts is not, in principle, privileged.

The problem with this line of reasoning is that it places too strict limitations one what we can and cannot be said to possess authority over. That is, if we accept that mental events are *not* physical events (that mental events are not in principle knowable by others), we restrict privileged access to essentially private knowledge claims: claims or thoughts we have to ourselves, but in no way express. We could not, given these assumptions, have first-person authority over thoughts we express to others. If we follow Davidson's understanding of first-person authority, however, there are many more possibilities for actual instances of (privileged) self-knowledge. For example, suppose I have a thought that involves the word 'food' and you do not speak my language. I can effectively communicate my thought about the word 'food' to you only if you presume that I know what I am talking about, I am consistent with my usage of the word 'food' (or whatever verbal representation of 'food' I employ while trying to convey the meaning of 'food' to you), and you are able to recognize my patterns of speech when I am speaking about 'food.' According to Davidson, I come to know what I mean by 'food' in a way

completely different from the way in which you come to know the meaning of 'food.' I have special epistemic authority over what I mean; your epistemic status with regard to 'food' is derived, or secondary, if you will. The point is simple: my privileged knowledge of the meaning of my word 'food,' or of my mental state or propositional attitude, etc. does not preclude you from also knowing those things. You simply know them differently than I do.

"Even so," the objector might insist, "Davidson's compatibilism rules out the possibility of first-person authority with respect to the meaning of his words!" Davidson's appeal the external determinant of the truth of an individual's statement is not the issue; the meaning of an individual's word is the issue. Regardless of whether one's statement is true, whatever meanings one applies to words, those meanings have been socially determined, either appropriately or inappropriately. If one does not learn the meaning of a word appropriately and accurately, e.g. the speakers in an individual's community mistakenly think 'arthritis' refers to a condition of the joints caused only by calcium deposits; this (false) meaning of 'arthritis' the individual endorses is still socially determined. The error is a result of a mistake the linguistic community has made about the meaning of 'arthritis.' Davidson wants to claim both (1) that the individual gives meaning to his words and (2) that the meanings of his words are socially determined. But (2) prevents (1). It is unclear, then, how an individual can have first-person authority over what he means by a word. Perhaps all Davidson is getting at is that the individual must only have first-person authority over the fact that he has a thought, but that brings us directly back to Burge.

CHAPTER 5 AKEEL BILGRAMI

5.1 BILGRAMI'S EXTERNALISM

So far I have presented arguments against Putnam's theory of meaning and reference and the externalism it engenders, as well as criticisms of Burge's and Davidson's compatibilist strategies with respect to externalism and self-knowledge. Now I will examine both Akeel Bilgrami's suggestion that Burge and Davidson have failed to reconcile self-knowledge with Putnam's externalism precisely because such an externalism necessarily precludes self-knowledge, and Bilgrami's new version of externalism, as found in "Can Externalism Be Reconciled With Self-Knowledge," that is not dependent upon Putnam's causal theory of meaning and reference, Bilgrami contends that not all externalist positions are so dependent and, moreover, that his own version of externalism averts all possible threats to self-knowledge. Let me spell this out more clearly.

The externalist thesis is generally defined as the denial of internalism, or what Putnam calls 'methodological solipsism': namely, it is the denial of the claim that intentional (or mental) states are independent of the existence of anything external to the individual who possesses those states. Bilgrami suggests that this denial is the minimum requirement for an externalist thesis and denotes any externalism that simply satisfies this requirement as general externalism, or (G.E.). Putnam's causal theory of meaning and reference claims that the meanings of one's concepts, which in turn determine the contents of one's thoughts, are determined by the scientific essences of the concepts' external referents. Though this externalist picture certainly entails (G.E.), it is *not* the case that (G.E.) entails Putnam's externalism. Not all externalist theses are committed to Putnam's causal theory of meaning and reference and *its* commitment to a specific notion of externalism, that meaning is dependent upon a God's-eye view of things in the world.

According to Bilgrami, most contemporary philosophers of mind and language who find anything "worthwhile in externalism take it for granted that externalism must be cashed out in terms of [Putnam's] views" regarding an item's microstructural properties as the ultimate determination of the meaning of concepts. (Bilgrami, 234)[1] Given the orthodox acceptance of Putnam's specific externalism, Bilgrami refers to Putnam's externalist picture as 'orthodox externalism,' or (O.E.). If we were to conflate (O.E.) with (G.E.), as so many philosophers have done and continue to do, it is easy to see why externalism apparently threatens self-knowledge so forcefully. That is, if we limit ourselves to Putnam's externalism, we must be willing to admit that if one (or one's community's expert) lacks knowledge of the external determinants of his thought contents, i.e. the 'scientific essences' of the things his thoughts are about, one must also lack knowledge of the contents of one's thoughts. This conclusion, as we have seen, is too counter-intuitive for many to accept.[2] However, says Bilgrami, there is no reason to believe (G.E.) commits us to (O.E.), and *its* denial of self-knowledge. What is needed, then, if we are to maintain our intuition that we do know what we think we know, is to define a different specific externalism that satisfies (G.E.) but does not threaten self-knowledge.

Bilgrami proposes a much different compatibilist strategy than that of Burge and Davidson, claiming that their strategies are destined to fail to save self-knowledge from the externalist threat because they fail to disarm the threat Putnam's externalism raises against self-knowledge. He contends that Putnam's externalism does not appropriately constrain the external determination of concepts and, therefore, thought contents. His compatibilism rests on redefining a specific externalist thesis in terms of a constrained notion of external determination that thereby averts all possible threats to self-knowledge.

---

[1] Bilgrami, "Can Externalism Be Reconciled With Self-Knowledge."
[2] Hence, we see the desire of Burge, Davidson, and Bilgrami, to name just a few, to reconcile self-knowledge with externalism.

Bilgrami describes the constraint on the external element as follows:

> (C): *When fixing an externally determined concept of an agent, one must do so by looking to indexically formulated utterances of the agent which express indexical contents containing that concept and then picking that external determinant of the concept which is in consonance with other contents that have been fixed for the agent.* (Bilgrami, 255)

When Bilgrami talks of "indexically formulated utterances" and "indexical contents" he is simply referring to those assertions the agent makes about things in the world. That is to say, if an agent's thought is about 'water' or 'arthritis,' his indexical utterance refers to some item in the world (the indexical) and therefore makes use of the indexical concept 'water' or 'arthritis.' The constraint requires that in fixing an agent's concepts not only must one pick out the particular item in the world that correlates with the term used to express the agent's content, one must also describe the external determinant of the concept "in a way that fits with the other *contents* one has attributed to the agent." (Bilgrami, 257) Thus, to fix the concept, say, 'clutch,' one must consider both whether there is some sort of pedal regularly present under the agent's foot while sitting in a car when he utters something about clutches and whether he has at least some minimal understanding of the 'inner workings' of automobiles.[3] The point of the second requirement entailed in the constraint is that there is no appeal to analyticity when 'fixing' the meaning of a concept. Since concepts are always to be mediated by the agent's beliefs and other concepts, there cannot, on Bilgrami's view, be one, analytic definition of any concept.

---

[3] Bilgrami notes a few caveats that must be established before his externalism can take root:

1) We must assume that meanings and concepts are public in that the agent does not have sole possession of their constitutive conditions.
2) 'Concept' is to be understood as the counterpart to 'term' just as 'content' is the counterpart to 'sentence.'
3) When Bilgrami talks of 'fixing' a concept (and therefore content) it is a way distinct from Putnam's 'direct reference.' The second aspect of his constraint requires making use of the agent's entire belief set in fixing meaning.
4) We must confine ourselves to sincere first-person utterances; lies and metaphors are unhelpful in fixing meaning.
5) Not all concepts have 'simple' external correlations, i.e. the concept of a unicorn has no actual external referent, but it can be fixed complexly: the concepts of a horse and a horn are combined to fix the meaning of the concept 'unicorn.' (Bilgrami, 255-56)

Picking out the proper correlating item in the world requires "a) shared similarity standards, presumably wired into us all, so that what is grossly salient to him is not wholly at odds with what is salient to us, and b) Mill's methods."[4] (Bilgrami, 257) Let me illustrate the application of a) and b) with a brief example. Suppose an agent says something like, "Water will quench my thirst." According to a), the substance the agent refers to as 'water' must be a substance that other agents can easily and readily 'pick out' in the environment. That is, the speaking agent cannot have an entirely unique capacity for picking that substance out of the environment. Furthermore, b) tells us that if we employ, say, Mill's method of agreement, we will be able to determine which substance 'water' refers to. In other words, if there is an array of substances in the agent's environment whenever he utters, "Water will quench my thirst," application of the method of agreement will allow us to determine which substance in the array the agent is referring to as 'water.' If substances *X*, *Y*, and *Z* are present in the agent's environment, but only *X* is *always* present when he makes the indexical utterance, "Water will quench my thirst," then *X* must be the substance the agent refers to as 'water.'[5]

However, the constraint suggests that merely picking the proper salient item in the world that correlates properly with the indexical utterance of an agent is not sufficient for providing a new externalist thesis about the determination of the meaning of concepts. If picking out the right object in the world as a concept's referent *were* sufficient for fixing the meaning of a

[4] Mill's methods serve to explain causal relations and, despite criticisms and modifications since Mill first explored them in *System of Logic*, they function implicitly in many ordinary inductive inferences. They are as follows:

1) Method of Agreement – identifies a sufficient condition for an event by finding a common factor;
2) Method of Difference – identifies a necessary condition for an event by finding a single deviate factor;
3) Joint of Method of Agreement and Difference – identifies a necessary condition for an event by finding a common factor that is both present among two or more occurrences and absent among two or more other occurrences;
4) Method of Residues – identifies an unknown causal connection by ruling out known causal connections;
5) Method of Concomitant Variation – matches variations in one condition with variations in another. (Hurley *A Concise Introduction to Logic*, pp. 487-497)

[5] I do not mean to suggest that *only* Mill's method of agreement can be used to pick out the indexical item in the world. In fact, some situations may arise in which a different method or a combination of several methods might be necessary for determining which substance the agent is making a reference to in his indexical utterance.

concept, this 'new' externalism would not be much different from Putnam's. What is peculiar to Bilgrami's externalism is the further requirement produced by the constraint. That is, (C) tells us that we must also describe the external determinant, i.e. the correlative item in the external world, not according to its microstructural properties, but in such a way that it "fits in with the other *contents* one has attributed to the agent." (Bilgrami, 257) Contents are attributed to agents only after concepts are attributed to them. That is, no agent can have, say, 'clutch-thoughts' unless he has the concept of a clutch, which in turn requires the agent to have other concepts, i.e. those of pedals, transmissions, etc. Let us look at Bilgrami's example for clarification of this.

Suppose there is a pedal regularly present under an agent's left foot whenever he utters indexical statements containing some term that refers to that pedal. According to the constraint, we cannot properly describe the external determinant, i.e. the pedal, as a 'clutch' unless we have reason to believe that the agent has some other beliefs about the inner workings of a car. That is, one could not attribute the concept of a clutch to the agent unless one also attributes to the agent the knowledge that the pedal served the purpose of being a clutch, so to speak.[6] Moreover, the agent's thought content would not be about a clutch if he did not possess the concept of 'clutch.' Bilgrami's point is that the agent's 'aggregate' beliefs must be considered when attributing a particular concept and thus particular thought content.

We can now begin to see why Bilgrami's version of externalism does not rely on a theory of direct reference and does not beg for analytic definitions. That is, the 'aggregative level' of content attribution is the set of all of the agent's beliefs about certain terms or concepts from which one can attribute specific meanings to those concepts the agent employs. All concepts, therefore, are mediated by the agent's arsenal of beliefs. Some concepts are obviously less

[6] Of course this talk of 'serving a purpose' might be misleading. Bilgrami is not suggesting that some teleology is involved in the external determination of concepts. Rather, his point is that the agent's aggregate beliefs must be considered when attributing a particular content to a particular belief.

mediated than others, i.e. the concept of a 'pedal' is less mediated than the concept of a 'clutch.' If we were to consider only the aggregate level of content attribution, many terms or concepts would be very 'fine-grained;' that is, different agents could have 'very numerous and very diverse' notions of 'clutch' or 'water' or 'arthritis.' In other words, since each agent is bound to have his own peculiar set of beliefs, employing only the aggregate level of content attribution would make it very difficult to establish a common meaning of a particular concept.

Therefore, Bilgrami contrasts the work that can be done at the 'aggregative level' of content attribution, i.e. the determination of meaning, with the work that is done at the 'local level' of content attribution, the explanation for why two agents who perform the same actions share the same concepts. The fact that there will be a large and diverse set of beliefs relevant to a concept plays no role in the attribution of specific content to explain behavior. Content, therefore, is determined only at the local level, e.g. the content of the agent's belief that water will quench his thirst explains his drinking the water. Any other thoughts the agent has are irrelevant to the particular action of drinking the water. An agent's chemical beliefs (or lack of chemical beliefs) about water do not factor in the action of drinking the water. This point is sticky, so once again I will use Bilgrami's own words.

> The reason why [the diverse set of beliefs employed at the aggregate level] does not matter is that these attributions of concepts are attributions of things that *do not go directly* into the attribution of specific contents to explain behavior. All that these attributions do is to provide a pool of resources which one uses in a selective way in order to attribute specific contents in the explanation of behavior. (Bilgrami, 259)

Consider Bilgrami's own example to make this as clear as possible.

Suppose one agent knows a fair amount of chemistry and another agent knows none. At the 'aggregative level' the two agents' concepts of 'water' are different, given their different beliefs about the chemical composition of water, i.e. one agent has such beliefs and the other

does not. "[The aggregate level] is the level at which theories of meaning do their work." (Bilgrami, 259) It is *not*, however, the level at which action explanation takes place. At the level in which action explanation *does* take place, the 'local level,' only those beliefs that are relevant to an agent's action play a role in determining the particular concept and, therefore, its corresponding content of the agent's thought. That is, both the chemically knowledgeable agent and the chemically ignorant agent will employ the same local concept of water when they drink it: presumably, they both believe that it will quench their respective thirsts. Locally determined concepts, and therefore content, account for the fact that we often do share concepts with other agents, despite the fact that those at the aggregate level seem to be 'fine-grained.'

Bilgrami's claim that all concepts are mediated by an agent's set of beliefs, which in turn determine his other concepts, appears to suggest that he has a foundationalist picture of content determination. That is, if all of an agent's concepts are mediated by the agent's beliefs, it seems, then, that at some point we must find the one, 'foundational' belief that gives rise to all of the agent's concepts. Bilgrami adamantly denies this charge, saying, "[There] is a strong element of anti-foundationalism built into this constraint." (Bilgrami, 257) He claims that *all* the agent's concepts are 'fixed' relative to other concepts the agent possesses, and therefore, none of the agent's concepts can serve as static foundations.[7] There can be no 'fixed' meaning of a concept, in the sense of being rigidly designated, because all concepts are mediated by each agent's belief set, which will undoubtedly produce a wide array of meanings. We should not conflate his anti-foundationalist position with respect to concept determination with the epistemic claim that no meanings can be fixed since the growth of knowledge entails constant revision of the meaning of concepts. Rather, his anti-foundationalism is meant to show that meanings cannot be analytic because each concept must be defined relative to the agent's other concepts.

[7] I will present an objection to Bilgrami's anti-foundationalist view in section 5.3.

One might argue further that Bilgrami is "insisting on an internalist filter upon the external" when he speaks of fixing an agent's concept relative to that agent's other concepts. (Bilgrami, 258) Again, Bilgrami denies such a charge, claiming that *all* the agent's concepts are externally determined. "There is, to begin with, something misleading, in fact downright false, in thinking of the filter as internal since the belief contents of an agent which provide the filter will contain concepts, which are themselves externally determined." (Bilgrami, 258) He thinks the criticism comes from mistakenly thinking that we only have two options: either endorse a direct externalism (i.e. Putnam's externalism) or assume internalism.[8] He holds that since his externalism satisfies both his constraint and (G.E.) it cannot possibly be an internalist thesis. This defense is somewhat suspicious, but I will defer my criticism to the last section of this chapter. Suffice it to say, for now, that Bilgrami leaves mysterious how any of the agent's concepts that are used to fix other concepts can be externally determined in the first place.

Bilgrami concludes that a specific account of externalism that follows his constraint and its commitment to the claim that content is determined *only* by local concepts will pose no threat to self-knowledge. That is, because his externalism insists on an external determinant that is mediated by the agent's beliefs, and does not invoke the 'objective natures' of things in the world that correspond to the concepts agents employ in expressing the content of their thoughts, self-knowledge is not threatened. To see this conclusion clearly, though, we must first understand his criticisms of Burge and Davidson. The next section of this chapter will examine those criticisms in such a way as to bring to bear what Bilgrami calls the 'indirect strategy' for posing the question of self-knowledge in light of Putnam's externalism. According to Bilgrami, this 'indirect strategy' provides insight into how neither Burge nor Davidson has appropriately reconciled self-knowledge with externalism.

[8] See section 5.2 for Bilgrami's reasons for claiming that this is a false dichotomy.

5.2 BILGRAMI'S CRITICISMS OF BURGE AND DAVIDSON

Since I have already said quite a bit about both Burge and Davidson, I will be brief with respect to Bilgrami's criticisms of their compatibilist strategies. Primarily, he claims neither Burge nor Davidson has appropriately liberated self-knowledge from the externalist threat simply because they have failed to aptly pose the question of self-knowledge. That is, they do not pose the question of self-knowledge is such a way as to bring to light the true threat Putnam's externalism begets. One reason for this is their failure to see that Putnam's externalism is not the only externalism that satisfies (G.E.), that we are not stuck with Putnam's scientific essentialist understanding of the external determination of meaning.

Recall Burge's compatibilist strategy. He claims that all that is needed to thwart the *prima facie* threat to self-knowledge, i.e. the claim that if mental states are constituted, at least in part, by things external to the agent, the agent may not know those states unless he knows the external factors that constitute them, is to demonstrate cases in which the agent does know his mental state (or, in other words, the content of his thought). Burge calls such cases *basic self-knowledge*, which show that since an agent is thinking *that p* when he knows that he is thinking *that p*, even if *that p* is externally constituted, those external constitutive conditions cannot in principle threaten self-knowledge. Basically, says Bilgrami, Burge's *basic self-knowledge* shows that one cannot argue directly from externalism to lack of self-knowledge.

However, Burge's compatibilism merely shows that the *prima facie* threat to self-knowledge is poorly posed. Suppose we can show that self-knowledge is indirectly threatened by Putnam's externalism. Bilgrami contends that Burge's compatibilist strategy cannot thwart *that* kind of threat. This 'indirect strategy' for posing the threat to self-knowledge brings to bear the full implications of Putnam's casual theory of meaning and reference and the externalism that

it produces. That is to say, if we assume that the meaning of a concept is determined by the microstructural properties of the concept's external referent, we must be willing to attribute inconsistent thoughts to agents in particular circumstances.

For example, suppose Joe has no clue as to the microstructural properties of water, or that he does not possess the medical experts' knowledge of arthritis. Now, suppose Joe claims, "I believe water is not H_2O," or "I believe I have arthritis in my thigh." According to Putnam's view of meaning, Joe must be saying something tantamount to the blatantly inconsistent claims of "I believe the substance with the chemical composition H_2O is not H_2O," or "I believe I have a disease of the joints only in my thigh." (Bilgrami, 240) Of course, given the principle of charity, we do not conflate lack of chemical or expert knowledge with 'logical idiocy;' that is, we do not generally think people often espouse sincere, yet blatantly inconsistent, first-person beliefs. To avoid the attribution of 'logical idiocy,' we might say that Joe need not know that his claims are tantamount to blatantly inconsistent claims. But to say *that*, we must also say, given Putnam's causal theory of meaning and reference, that Joe does not know the meaning of the words (concepts) he employs in his sincere first-person utterances, and so does not know what he thinks he knows.

Self-knowledge, then, is threatened by an indirectly induced dilemma: either attribute blatantly inconsistent beliefs to the agent, or attribute lack of self-knowledge to the agent. Putnam suggests that the only way out of this dilemma is to bifurcate content into two notions: externally determined (wide) content and internally determined (narrow) content. This compatibilist move by Putnam suggests that he is unsettled by the apparent consequence of his externalism – that self-knowledge is not possible, at least not with respect to those concepts

whose external referents are defined by their 'scientific essences.' Burge is unwilling to accept this bifurcation, so we might take this approach as the third horn of a trilemma.[9]

Bilgrami poses three possible Burgean responses to his charge that the basic self-knowledge compatibilist strategy is not enough to disarm the specific threat Putnam's externalism poses to self-knowledge as raised by the 'indirect strategy.' First, the Burgean compatibilist may respond by charging that it is unfair to 'rewrite' Joe's claims about water and arthritis in terms of things of which Joe is, *ex hypothesi*, ignorant, i.e. that 'water' is the substance with the chemical composition H_2O and that 'arthritis' is a disease of the joints only. Burge, himself, suggests such a response when he makes a distinction between 'concept' in the sense of the lexical item in the world (the external referent) and an agent's 'conceptual explication,' what the agent would give as his understanding of the meaning of the concept he employs. According to Burge's social externalism, the agent's 'conceptual explication' need not be the same as an expert's 'conceptual explication' in order for the agent to know what he thinks. Bilgrami claims that this response 'surreptitiously concedes' that there are two notions of concepts, and by implication, two notions of content: "a) concepts proper, given by external reference and b) concepts in the sense of conceptual explications the agent can, on reflection, articulate." (Bilgrami, 242) Thus, this Burgean response avoids the first horn of the trilemma (of attributing inconsistent beliefs to the agent) by "impaling [Burge] on the third horn" (bifurcating content). (Bilgrami, 243)

The second proposed Burgean response is that Bilgrami's 'rewrites' or substitutions of the meaning of the concepts Joe employs with that which is specified by the objective natures of 'water' or 'arthritis' implies something like the analytic-synthetic distinction, which Burge and

[9] Time and space limitations require me to accept without argument Bilgrami's claim that Burge, in "Individualism and Psychology," views bifurcation as an inappropriate option.

Putnam have both denied. Bilgrami replies that if such rewrites are not relevant, what could we possibly salvage from Putnam's externalism? That is, such rewrites bring to light exactly what Putnam's causal theory of meaning and reference is designed to produce, namely, the extension of any word or term. If we accept the second response, we must either abandon Putnam's externalism or offer an alternative account of what Joe's concepts of 'water' or 'arthritis' could possibly mean.

Such an alternative rewrite might employ a *metalinguistic specification* of what Joe means by his concepts. That is, we might simply say that Joe believes that 'water' and 'arthritis' mean whatever the experts say they mean. On this account of Joe's beliefs, we have retained the (unspecified) relevance of scientific essences in the determination of meaning. But, Bilgrami retorts, now the external element enters in the meaning of the agent's concepts mediated by the agent's beliefs. Surely the agent's belief concerning the meaning of 'arthritis' is different than the expert's belief about that meaning. "The [external] reference is no longer crucial in the specification of concepts, it is the *differing* beliefs or descriptions of the relied upon [the expert] and the relying agent [Joe] which are doing the work, so the concepts attributed to them will be quite different." (Bilgrami, 244) We are no longer considering an externalism that is based directly on Putnam's causal theory of meaning and reference. Since Burge set out to reconcile Putnam's externalism with self-knowledge, which this alternative account of Joe's beliefs does not endorse, the only way for Burge's compatibilism to work is to make use of a different externalism than Putnam's.

Bilgrami's reply to the third proposed Burgean response illuminates what I believe to be central to the new externalism Bilgrami posits. The Burgean could respond to the proposed rewrite by claiming that there is no need for a rewrite. All that is needed is the concept of

'water' or 'arthritis.' That is, according to the Burgean position, there is nothing wrong with the fact that Joe may falsely believe that he has arthritis in his thigh. Joe's mistaken notion of the meaning of the concept 'arthritis' is not enough to threaten his first person authority over his own (false) thought or belief. Bilgrami takes this response as simply and mysteriously claiming, "'Arthritis' refers to arthritis." This response is mysterious because it begs the question: what does the disquotational term mean? The Burgean compatibilist, then, might say that Bilgrami is insisting on definitions.

Bilgrami says that it is quite wrong to suspect the line of questioning regarding the meaning of the disquoted term in the Burgean's third response as insisting on definitions. "Disquotation, if it is to be in the service of an account of meaning, is not a wholly trivial idea. It must be anchored in something which is not made explicit in the disquotational clause itself." (Bilgrami, 246) That is, the disquotational clause must express something if the response is to make any non-trivial sense at all. But the only thing it could express, on Putnam's externalism, is the "inconsistency-inducing 'a disease of the joints only,' since that is what the scientific experts think arthritis is." (Bilgrami, 246)

One could say that the disquoted term refers to some object in the world that cannot be elaborated in definitive terms. That is, we might not take the disquotational strategy as merely a syntactic device, but rather as suggesting some 'metaphysical hook up' between the meaning of the concept 'arthritis' and some object in the world; that there is some causal relation "unmediated by any description" that yields the meaning of the concept 'arthritis.' Surely some concepts are primitive in the sense that they may not be susceptible to rewritable explication. After all, explications are always given in terms of other concepts, which invokes an infinite regress of explications if there are *not* some primitive concepts.

The appeal to disquotational assertions and the primitiveness of some concepts yields the conclusion that at least some concepts are in principle inexpressible. Bilgrami finds this conclusion quite unsatisfying because it leaves semantics 'ineffable and mysterious.' That is, if concepts are, by their nature, inexpressible, then their meaning must be wholly mysterious. The Burgean compatibilist who endorses this third response is committed to this mysteriousness. Those who endorse Bilgrami's version of externalism are not so committed because his "insistence on beliefs or descriptions being brought in to answer the question, [what does the concept *arthritis* mean?], precisely eschews this mysteriousness." (Bilgrami, 249) There is no mystery on Bilgrami's view because of his contention that all concepts are, in principle, expressible. Though he opens the door to an "infinitely regressive appeal to descriptions," he believes that routine pragmatism will overcome this difficulty, which is inherent in any anti-foundationalist view.

Consider Davidson now. Davidson's compatibilist strategy is equally susceptible to the 'indirect strategy' for posing the threat to self-knowledge. Recall that Davidson argues that the reason Putnam is committed to denying self-knowledge is because Putnam shares the internalist assumption that only objects of thought can account for self-knowledge. Since Putnam believes that there are no such objects of thoughts, he must conclude, argues Davidson, that one may not know the contents of his thoughts. Bilgrami, however, contends that Davidson's criticism of Putnam is off the mark. That is, according to Bilgrami, the assertion about the role objects of thought play in whether an agent can have self-knowledge is irrelevant to Putnam's conclusion that self-knowledge is not possible. Rather, Putnam's commitment to his 'scientific essentialist' view of the external determination of meaning is the sole reason he must deny self-knowledge.

Furthermore, Putnam does not deny all forms of self-knowledge, as Davidson falsely charges; Putnam denies self-knowledge only of those cases involving natural kinds.

Moreover, Davidson's positive remarks as to why we do, in fact, have self-knowledge "answer a question that has nothing specifically to do with Putnam's externalism." (Bilgrami, 253) That is, at the beginning of Davidson's article, he asked how self-knowledge could be compatible with Putnam's externalism.[10] Instead of answering this question, his thesis regarding the nature of interpretability answers quite a different question. Namely, it answers: "what, in general, explains the undeniable fact that agents whom we are interpreting by and large have *non-inferential* self-knowledge (first-person authority) of their own thoughts, given that in our interpretations we are not specifying objects of thought within their epistemological ken, but looking instead to external objects in their environment?" (Bilgrami, 253)

Davidson's positive explanation of first-person authority makes no mention of the interpreter's appeal to the objective natures of natural kinds, which is what prompted the question of self-knowledge in the first place. That is, Putnam's externalism suggests that lack of knowledge of the objective natures (or scientific essences or microstructural properties, whatever you prefer to call them) is the source of the threat to self-knowledge. Davidson fails to pose the threat to self-knowledge in *this* manner and, accordingly, is not successful in his attempt to reconcile self-knowledge with Putnam's externalism. In other words, Davidson's compatibilist strategy does help disarm the *prima facie* threat to self-knowledge; it shows that it does not follow directly from the fact that what we think is partly constituted by things that are external to us, that we do not know what we think we know. However, we have already seen via Bilgrami's criticisms of Burge that the *prima facie* threat to self-knowledge is poorly posed; that is, it does

[10] Davidson, "Knowing One's Own Mind," p. 92, claims that the "thesis of this paper is that there is no reason to suppose that ordinary mental states do not satisfy both conditions (I) and (II)" as found in Putnam's representation of the traditional theory of meaning.

not account for the specific threat levied against self-knowledge raised by Putnam's specific externalism. Neither Burge's nor Davidson's compatibilism can account for the indirect threat.

5.3 TWO CRITICISMS OF BILGRAMI'S NEW EXTERNMALISM

The specific version of externalism Bilgrami offers is not entirely satisfying for at least two reasons, both of which follow from his anti-foundationalism. My criticisms of Bilgrami are not simply criticisms of anti-foundationalism in general, however. That is, as an epistemic thesis there may be nothing in principle wrong with anti-foundationalism. The growth of scientific knowledge, for example, suggests that any particular definition or theory is susceptible to revision. The issue with which Bilgrami is concerned, however, is not an epistemic question; rather, he is (as are Putnam and Burge) concerned with the question of how the meanings of concepts are determined. My objections to Bilgrami's specific externalist thesis, therefore, ought not be construed as a hard-line foundationalist attack on anti-foundationalism since such an attack is more aptly applied to the epistemic question of whether the ultimate meaning of any concept can be known at all. I will show (1) that if Bilgrami's thesis is indeed anti-foundationalist, then the external element enters mysteriously into the determination of meaning, and (2) that Bilgrami's specific externalism commits him to a bifurcated notion of content.

My first criticism of Bilgrami's specific externalism follows directly from his claim that he is not committed to analyticity: that, as a result of following his constraint, there is no analytic definition of a concept's meaning.[11] Bilgrami contends that his externalism avoids all threats to self-knowledge because he constrains the external element of meaning determination by forcing the meaning of an employed concept to be in consonance with the agent's other beliefs and concepts. One might, then, charge Bilgrami with insisting on an internalist filter on the external.

[11] "I am claiming that the insistence on the rewrite (an insistence forced by the mystery attaching to any view that denies it) *by itself* does not commit one to the analytic-synthetic distinction. It is only if one combines the insistence with certain accounts of concepts or the meaning of terms that one is committed to analyticity." (Bilgrami, 263)

That is, if the meaning of a concept an agent employs is determined by the other concepts the agent holds, it seems that a concept means whatever the agent thinks it means. For example, suppose I say, "If I drink water it will quench my thirst." At my aggregate level I hold various beliefs about water, i.e. that it is wet, it is used for cleaning and washing, that drinking it will quench my thirst, etc. But I do not have any beliefs about the substance's chemical properties. My action of drinking the water is explained, says Bilgrami, at the local level in terms of one or more of my aggregate beliefs. It seems, then, that *I* determine the meaning of my concept 'water' according to my own set of beliefs, which is exactly what the internalist thesis claims.

As we have seen, however, Bilgrami claims that his externalism is not committed to analyticity and dismisses the charge that his two levels of concept determination amount to having 'an internalist filter' on the external element since *all* the concepts available to the agent at the aggregate level are externally determined. He claims: "But the whole point of distinguishing between the aggregative, meaning-theoretic level and the local level was to allow that there can be lots of different localities at the local level … There is, therefore, no definition." (Bilgrami, 263) It is unclear to me, however, how the meanings of the agent's aggregate concepts could be 'fixed' in the first place. Let me explain.[12]

Recall that the charge of analyticity was made against Bilgrami's insistence that disquotational assertions must be more than a mere syntactic device, that they must mean something. That is, he claims that responding to the question, "What does arthritis mean?" by saying nothing more than, "'Arthritis' means arthritis," amounts to saying nothing if the right-hand side of the assertion, the disquoted term, cannot be given a rewritten explication. He avoids the charge of insisting on analyticity by suggesting that one can give a rewritten explication in terms of one's aggregate beliefs. One is not forced into giving an analytic definition simply

[12] Sarkar, "Three Counter-Arguments: A Very Rough Draft"

because of the demand for something more useful than, "'Arthritis' means arthritis." Let us look closely, though, at what follows from Bilgrami's method of meaning determination.

Suppose there are a finite number of concepts in any language, *L*, that an agent can possess, say, concepts *A*, *B*, *C*, ..., *Z*. According to Bilgrami's theory of meaning determination, concept *A* must be defined in terms of concept *B*, concept *B* in terms of concept *C*, and so on until we arrive at concept *Z*. Now, we are faced with a trilemma: either concept *Z* is the foundational belief upon which all other beliefs are determined, i.e. *Z* has an ostensive definition via an indexical and is *not* given in terms of other beliefs; or *Z* must be defined in terms of another concept in the set, say, concept *A*; or our assumption that there is a finite number of concepts in any language is false. We can dismiss the third horn of the trilemma because having an infinitely regressive notion of the meaning of a concept yields a more mysterious notion of meaning determination than that with which we began. The second horn, which Bilgrami would have us endorse, is equally unhelpful in meaning determination since it yields only circular meanings for concepts. For example: "What does *X* mean?" "*X* means *Y*." "What does *Y* mean?" "*Y* means *X*." Finally, we are stuck with the first horn: meanings are nothing more than ostensive definitions; but this was the conclusion Bilgrami's externalism was supposed to avoid. We must conclude, then, if we are to accept that Bilgrami is not begging for analytic definitions, that Bilgrami's constrained external element enters mysteriously into the determination of meaning.

Let me raise the issue one more time. If Bilgrami is as opposed to analyticity as he says, how might he respond to the following three cases? *Case 1*: Suppose there is a concept *X*, such that it is associated at the local level of content with other concepts like, *P, Q, R, ..., Z.* Further, suppose that no agent will regard an object as *X* if it fails to have any property *P, Q, R, ..., Z.* It

seems that the term *X* is indeed analytically tied to the other concepts *P, Q, R, ..., Z.* That is, since in the absence of *P, Q, R, ..., Z*, *X* is not present, there appears to be great reason to think the meaning of *X must* be given an ostensive definition in terms of *P, Q, R, ..., Z.* Bilgrami's example of the two agents, one chemically knowledgeable of the chemical properties of water, the other ignorant of such properties, suggests that if both agents mean the same thing when they utter, "Water will quench my thirst," they must do so *accidentally* since any common local content cannot be tied analytically to the term 'water' they employ in their utterances. That is, it is unclear, given Bilgrami's externalism, whether the two agents could think *X* is water if *X* failed to have among its properties, *P, Q, R, ..., Z*, the property of quenching one's thirst, since Bilgrami denies such an analytic definition of 'water.' This leads us to the next case.

Case 2: Suppose there is a concept, *X*, such that it is associated at the local level of content with the concepts, *P, Q, R, ..., Z.* Further, suppose that it is *not* the case that no agent will regard an object as *X* if it fails to have any property *P, Q, R, ..., Z.* That is, *X*'s being associated with properties *P, Q, R, ..., Z* is simply a contingent fact. Bilgrami's view suggests that this must be the case if *X* has no analytic definition. Certainly, in this case, the agent's local content does determine the meaning of *X*, but such a case appears to be extremely rare. (It seems that it is true only in a remote possible world and not in the actual world.) Its rarity, then, does not give Bilgrami's externalism the support he desires.

Case 3: Suppose there is a concept, *X*, such that it is associated at the local level of content with the concepts, *P, Q, R, ..., Z.* Further, however, suppose that it is *not* the case that *all* agents will not regard an object as *X* if it fails to have any property *P, Q, R, ..., Z.* That is, some agents will associate properties *P, Q, R, ..., Z* with the concept *X*; other agents will associate properties *A, B, C, ..., O* with the concept *X*. The meaning of *X*, then, can only be accounted for

by a 'cluster theory' of meaning: that satisfying a certain set, or cluster, of properties is sufficient for determining the meaning of a concept. Bilgrami, however, expressly disavows such a cluster theory.[13] So, he must be willing to accept that each individual will give a different meaning to the concept *X*. As we have seen, he *is* so willing, since he claims that the point of distinguishing between the aggregate and local level is to allow for a diverse set of localities (meanings) at the local level. He has not, however, explained how the external element has entered. That is, it seems he has allowed for first person authority, since each agent will give *X* whatever meaning the agent wants to give, but he has not explained the compatibility of this result with the external element. It is still mysterious how the concepts *P, Q, R, ..., Z* or *A, B, C, ..., O* are externally determined in the first place.

This brings me to my second criticism of Bilgrami's compatibilist strategy. Recall that Putnam believes that the only way to save self-knowledge of contents involving natural-kind terms from the externalist threat is to bifurcate content into two notions: one internal, the other external. He does not, however, give an adequate explanation of how internal content can work in concert with the external notion of content, given that he expressly denies internalism. If we are to salvage anything from Bilgrami's compatibilist strategy, we must, I think, take his distinction between the aggregate level and the local level of content determination as nothing more than an explanation of Putnam's bifurcation of content. This is a serious objection to Bilgrami since he set out to reconcile externalism with self-knowledge without falling prey to Putnam's bifurcation of content. I am not suggesting that there is something inherently wrong

[13] "The so-called 'cluster' version of the descriptive theory of terms was an early response to a roughly similar charge made in a slightly different setting. My response is quite different. My response makes vital use of what I just called the thesis of the locality of content, which is an essential aspect of the overall externalist conception of intentionality that I am offering as an alternative to orthodox externalism." (Bilgrami, 262)

with bifurcating content. Rather, Bilgrami's compatibilist strategy denies using a bifurcation while making explicit use of it.

My first objection to Bilgrami showed that the analytic-synthetic distinction, he so adamantly denies, is hard to get rid of. If anything is to come from the notion that local concepts, and, therefore, contents, are fixed relative to the agent's aggregate concepts, we must endorse the first horn of the above trilemma: that *some* concepts are primitive in the sense of not being dependent upon other concepts for their meaning. More to the point, though, Bilgrami's contention that he has not bifurcated content into two notions seems *prima facie* false. That is, *if* we endorse Bilgrami's theory of meaning determination and accept that constraining the external element as Bilgrami suggests *somehow* lets externalism into the debate (and I stress *if*), then we must think of content in terms of two notions: aggregate and local. Though he masks this bifurcation in terms of two levels of concept determination, I see no reason to think that he has *not*, in fact, employed Putnam's own strategy of reconciliation. On a positive note, Bilgrami's explanation of the different work that is done at the aggregate and local levels does seem to give Putnam's strategy more strength. That is, Putnam merely suggests that having an internal component in the determination of meaning will help salvage self-knowledge of contents informed by natural-kind terms, but he does not provide an explanation of how that component could work in concert with the external component. Bilgrami appears to have provided a nice explanation of what Putnam simply hinted.

Consider an example. Suppose I have a variety of beliefs about water at the aggregative level, i.e. I believe that water is found in the lakes, rivers, and oceans in my environment, that it is used to irrigate fields, wash cars, and, when dammed, produces electricity, that if I am thirsty drinking water will quench my thirst and, finally, that it has the chemical composition H_2O.

Now, suppose my twin has the same beliefs about water at *his* aggregative level, except he believes the chemical composition of water is XYZ. Both my twin and I are completely identical in all other regards: our mental and physical states are the same, whenever I utter claims about water, so utters my twin. According to Putnam, when my twin and I utter, "Water will quench my thirst," either we mean something different or we do not know what 'water' means. Moreover, in order to preserve self-knowledge of the content of our thought that water will quench our thirst, there must be two notions of content: one that explains the external determination of the meaning of 'water' and one that explains how we know non-inferentially what we mean by 'water.'

Bilgrami's specific externalism based on his theory of meaning determination saves us from having to deny self-knowledge in this instance, not because it is free from Putnam's externalism (though it is), but because it explains the two notions of content found in Putnam's bifurcation. That is, my twin and I know the content of our thought because that content was determined in consonance with our other beliefs and concepts. The content 'water will quench my thirst' is determined by the belief we hold at the aggregate level concerning *that* property of water. The chemical composition of water was not invoked in determining what we mean by 'water' in our assertion that it quenches thirst. So, Bilgrami's theory of meaning is not the same as Putnam's; yet, Bilgrami does not reconcile self-knowledge with externalism in the way in which he set out to do.

That is, Bilgrami claims:

> But Burge and Davidson eschew this bifurcation because they think it arises, in part, from an unnecessary surrender in the face of the problem raised only *prima facie* by externalism for self-knowledge … The eventual point of my criticism will not be that all externalists are stuck with Putnam's bifurcated conception of intentionality. My solution, as I said, will rather be that we need to abandon orthodox externalism (O.E.), and fashion a new and alternative kind of specific

> position which satisfies (G.E.) but for which the problem regarding self-knowledge does not arise even *prima facie*. (Bilgrami, 236)

In other words, Bilgrami claims that not only are externalists not committed to a bifurcation of content if they are to save self-knowledge, they are not committed to Putnam's specific externalism. I have shown that though Bilgrami's specific externalism may be different than Putnam's specific externalism, it presents a dilemma for Bilgrami himself: either accept his two levels of concept determination, which suggests that even *his* compatibilist strategy is committed to a bifurcated notion of content *and* leaves mysterious the role of the external element at the aggregative level, or modify his externalism such that it admits of analyticity in order to eschew this mystery. As we have seen, Bilgrami does not want to follow either of these two paths.

CHAPTER 6 CONCLUSION

The aim of this thesis has been to examine various attempts at disarming the externalist threat to self-knowledge. One might conclude that the various objections I have posed to those who consider themselves to have disarmed this threat are evidence that the threat is an insuperable one. That is to say, though I stated at the outset that I am among those unwilling to accept that the denial of internalism entails a denial of self-knowledge, the structure of the previous four chapters, which shows that even the most prominent compatibilist strategies are wrought with their own inherent difficulties, suggests that I think externalism and self-knowledge cannot be reconciled. Such a conclusion would be quite wrong. I do admit that I have serious reservations about the compatibilist strategies presented above; however, I do not think that *that* entails the skeptical conclusion that there is no hope for reconciliation.

Indeed, any theory that claims to be a more successful account of the compatibilist position must address the concerns I have raised, and undoubtedly many others I have not. But the simple fact that no one, to my knowledge, has yet produced such a compatibilist theory does not further suggest that such a theory is, in principle, impossible. Rather, it may be the case that a truly successful compatibilism will be one that makes use of the successful components of each currently available compatibilist strategy. In what follows, I will give an extremely sketchy account of what ought to be employed in such an account. I do not, of course, presume that the following sketch is free from objection or even (dare I say) apparent inconsistency. It is, however, a useful starting point for putting the pieces back together.

First of all, I think a successful compatibilism ought to make great use of Putnam's views on the role of the division of linguistic labor and Burge's social externalist thesis. This strategy is quite dependent on the notion of epistemic reliabilism, which states that the agent need not

know all the conditions that must obtain for a certain event to occur. It should be enough that experts, both in the hard and social sciences, produce certain definitional meanings of concepts based on relevant environmental factors. If we employ the division of linguistic labor in conjunction with social externalism, we will be less disposed to limit external determination simply to natural-kind terms. Moreover, we will lessen the demands made upon the agent to know what his words and concepts mean.

Second, I think Burge's and Davidson's descriptions of first-person authority or privileged access are more or less accurate. As I have said before, their accounts are not entirely free from objection, but I do think they were on to something. Any successful compatibilism, then, ought to employ the idea that there is something inherent about first-person assertions that allows the agent to know what he is thinking in a direct and authoritative manner. However, it is not enough for a complete compatibilist account to simply claim that this is in fact how things are; it must also show how this aspect can work in concert with the external component.

As we have seen, Bilgrami's discussion of the aggregative and local levels of concept determination addresses how such a fusion of the internal and external might operate. What is lacking in Bilgrami's account is how the external component enters in the first place. Therefore, any successful strategy for reconciling self-knowledge with externalism must provide a convincing explanation of how the external element takes root at the aggregative level, but in such a way as to preserve the role of the local level in concept determination.

I do not believe that it will be an easy task to produce a compatibilist theory that satisfies all the conditions I have mentioned above. Nor do I think that these are necessarily the only conditions that must be satisfied. However, one should not conclude that difficulty amounts to impossibility. It is never easy to reconcile two seemingly conflicting intuitions!

BIBLIOGRAPHY

Bernecker, Sven. "Externalism and the Attitudinal Component of Self-Knowledge." In *Knowledge*, 499-511. Edited by Sven Bernecker and Fred Dretske. New York: Oxford University Press, 2000

Bilgrami, Akeel. "Can Externalism Be Reconciled with Self-Knowledge?" *Philosophical Topics* 20 (1992): 233-67

Boghossian, Paul A. (1989) "Content and Self-Knowledge." In *Knowledge*, 480-498. Edited by Sven Bernecker and Fred Dretske. New York: Oxford University Press, 2000

Burge, Tyler. "Individualism and Psychology." *Philosophical Review* 95 (January 1986): 3-45

--------. "Individualism and Self-Knowledge." In *Knowledge*, 468-479. Edited by Sven Bernecker and Fred Dretske. New York: Oxford University Press, 2000

Davidson, Donald. "First-Person Authority." *Dialectica* 30 (1984): 101-112

--------. "Knowing One's Own Mind." In *Externalism and Self-Knowledge*. 87-110. Edited by Peter Ludlow and Norah Martin. Stanford: CSLI Publications, 1998

Hurley, Patrick. *A Concise Introduction to Logic*. Belmont: Wadsworth 2003, 487-497

Putnam, Hilary. "The Meaning of Meaning." In *Externalism and Self-Knowledge*, 87-110. Edited by Peter Ludlow and Norah Martin. New York: CSLI Publications, 1998

VITA

Gabriel Guy Cate is a Graduate Assistant in the Department of Philosophy and Religious Studies at Louisiana State University in Baton Rouge. He plans to graduate in Spring 2003 and hopes to attend a doctoral program in philosophy in the future. He is originally from the Austin, Texas, area but attended high school in Shreveport, Louisiana, where he met his future wife, Allison Yvette Seeliger. The two were married in 2001. When not studying philosophy, Gabriel enjoys playing his guitar and writing music.