JUSTIFICATION INTERNALISM, SELF KNOWLEDGE, AND MENTAL CONTENT EXTERNALISM

By
Amber Ross

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Philosophy.

Chapel Hill
2006

Approved by:

Advisor: Ram Neta

Reader: Marc Lange

Reader: John Roberts

ABSTRACT

Amber Ross: Justification Internalism, Self Knowledge, and Mental Content Externalism
(Under the direction of Ram Neta)

At first blush, mental content externalism and justification internalism seem incompatible. If some of the content of my mental states supervenes on factors external to me, the content of these mental states might be unavailable to me. If the factors relevant to the justification of my beliefs are the relations between the contents of my beliefs, and I do not have access to these contents, then these beliefs cannot be justified internally.

I propose to reconcile mental content externalism with justification internalism by taking the factors relevant to the justification of a belief to be the relations between how one would express one's beliefs, not between the contents of those beliefs. Though mental content externalism may somewhat restrict an agent's self knowledge, it could not restrict an agent from knowing how he would express his beliefs, and therefore would not hinder his access to the relevant justificatory factors.

TABLE OF CONTENTS

I Introduction

Justification internalism is the view that all the factors relevant to the justification of a belief are in some sense internal to the agent and available to her. Mental content externalism is the view that the content of certain types of intentional mental states is in part determined by factors outside the agent. There has been expansive debate over whether justification internalism is a satisfactory theory of justification, or instead if some factors external to the agent, and not necessarily accessible to her, determine whether a belief is justified. More recently, there has been a related debate concerning whether mental content externalism, a theory that enjoys broad acceptance in philosophy of mind, poses a special problem to justification internalism. The problem would be this: If some of the content of my thoughts supervenes on factors external to me, this aspect of my thought might be unavailable to me. Mental content externalism would thereby place restrictions on my self knowledge. The justification internalist maintains that the factors relevant to the justification of my beliefs are factors to which I have first-personal access, factors that I can know from my first-person perspective. If the factors relevant to the justification of my belief are the relations between the contents of my thoughts to which I do *not* have access, my beliefs could not be justified internally.

In this paper, I will attempt to provide an account of self-knowledge, as restricted by mental content externalism, that would put the factors relevant to justification within the

agent's first person perspective and would thereby be accessible to the agent. If my account is accurate, then we will have found a way in which justification internalism and mental content externalism would be compatible, even if we accept all the limitations mental content externalism might place on self-knowledge. In doing so I will discuss forms of justification internalism characterized by Conee and Feldman and by BonJour, the account of mental content externalism given by Putnam and Burge, the debate between Boghossian and Burge on self knowledge, and James Chase's reply to BonJour's speculation that mental content externalism and justification internalism are incompatible.

II The Views

Justification Internalism

As epistemic agents, our first desire may be that we hold only, or at least mostly, *true* beliefs both about ourselves and about the world around us and we would prefer not to hold these true beliefs merely by luck. We also want to rightly feel *confident* that our beliefs are mostly true, confident that the way in which we form our beliefs will continue to be trustworthy. We hope that we form our beliefs *rationally* and that we refrain from believing things that we have no good reason to believe. So long as I am rational in forming my beliefs and hold them for good reasons, I will be satisfied that my beliefs are justified and as confident as I can be that I am rational, whether or not my beliefs are true. These are at least some of our desiderata as epistemic agents.

Justification internalism is the view that an agent's beliefs are justified if that agent has good reasons, internal to the agent herself and (according to most internalist positions) to which she has access, for holding the beliefs that she does. Of course, there are a variety of internalist positions, but one element they share in common is that they all seem to use "justification" in a way that fits with our pre-theoretical intuitions regarding what it means for our beliefs to be justified. An agent's belief is justified if that agent had no way of knowing (or no reasonable way of becoming aware of the fact) that this belief was false or

based on dubious reasoning. Justification internalism is attractive because it puts all the factors required for judging oneself to be a reasonable epistemic agent within one's. It allows the agent to be fully responsible for the justification of her beliefs.

If we adopt justification internalism, justified beliefs will fit the desiderata outlined above. On this view, for my belief to be justified I must have been rational in forming my belief and hold it for good reasons, reasons that are accessible to me from my current first-person perspective. Thus I will be satisfied that I hold the beliefs I ought to hold and as confident as I can be that I am rational, even though the actual truth or falsity of my beliefs is a matter outside my control.

Mental Content Externalism

Mental content externalism is the view that the content of certain types of intentional mental states is in part determined by an agent's relationship with the environment, and therefore does not supervene on physical properties within that agent. This theory of mental content was motivated by semantic content externalism, the view that the meaning and reference of some kinds of terms is determined in part by factors external to the agent. The most widely recognized arguments for semantic externalism, as well as mental content externalism, are Putnam's "Twin Earth" thought experiments.

Twin Earth is a planet on which there is an odorless, tasteless liquid that fill the rivers and oceans, expand when frozen, etc., which is superficially identical to water on earth. Also on Twin Earth, there is a community that speaks English, and they call this liquid 'water'.

However, the chemical composition of twin water (the water on Twin Earth) is XYZ. Aside from their chemical composition, properties of twin water and water are identical, and we might add that every other aspect of Twin Earth is identical to Earth as well.

Putnam argues that when Twin Earthlings use the word 'water', they refer to the liquid in their environment that goes by this name, XYZ. When Earthlings use the word 'water' they refer to the nearly identical liquid in their environment, H_2O. Therefore, following Putnam, we would judge that "On Twin Earth the word 'water' means XYZ" (1975, p585), and we would also judge that "On Earth the word 'water' means H_2O".

To explain the difference between the meaning of 'water' on Earth and 'water' on Twin Earth we need to look not to the internal states of the English and Twin English speakers, but to the environments in which each are immersed. What determines that 'water' means H2O on earth instead of meaning XYZ is that the earth environment contains H2O, not XYZ. The meaning of 'water' on Earth, therefore, is partially determined by the environment of the English speakers, partially determined by the actual extension of the term. And, of course, the same holds for the meaning of 'water' on Twin Earth.

It is a short leap from semantic externalism to mental content externalism. When a Twin English speaker entertains the proposition expressed by the sentence "There is a glass of water on the table," mental content externalists contend that he thinks a different thought than an English speaker whose thought would be expressed as "There is a glass of water on the table". Putnam claims that even if English speakers and Twin English speakers have not yet discovered the chemical composition of water and twin water, and are in the same

psychological state[1], they understand the term 'water' differently (1975, p585), and therefore the content of their mental states would not be identical.

In "Individualism and the Mental", Tyler Burge introduces a thought experiment similar to Putnam's Twin Earth involving a patient's understanding of the condition of arthritis and the social environments in which the patient might be immersed. As with Putnam, Burge concludes that due to differences in the linguistic community in the actual and counterfactual situations he describes, the agent's mental content involving the term 'arthritis' will be different in the different contexts, even though "the patient's internal qualitative experiences, his physiological states and events, his behaviorally described stimuli and responses… remain constant, while his attitude contents differ". (1979, p601)

In Burge's thought experiment, an agent S has a certain set of beliefs about arthritis that are based on "casual conversation or reading, and never hearing anything to prejudice him for or against applying [the term "'arthritis'] in the way that he does" (1979, p600). In particular, S correctly believes truly that he has arthritis in his joints, that he has had this condition for many years, etc. He also "thinks falsely that he has developed arthritis in the thigh" (1979, p600).

The reason that his belief is false, of course, is that the experts in his linguistic community delineated a set of medical conditions that are the extension of the term 'arthritis', and no condition of the thigh is a member of the set. That is, given the correct usage of 'arthritis' in his social environment, arthritis cannot be a condition of the thigh.

In a counterfactual linguistic community, 'arthritis' is taken to apply to all the conditions that it encompasses in the actual linguistic community, plus conditions of the thigh. Call the

[1] Assuming that we take "psychological state" here to supervene on properties internal to the agent, which is standard but not necessarily completely uncontrovertial.

agent in the counterfactual situation S*. S* would assent to all the same propositions that he would in the actual situation, and in the counterfactual situation his belief that he has arthritis in the thigh is correct. The question before us, then, is whether the content of S's mental states in the context of the actual social environment is the same as the content of S*'s mental states in the counterfactual environment.

Burge concludes, of course, that the content of S's and S*'s mental states are not identical.

> The word "arthritis" [in the counterfactual case] does not mean *arthritis*. It does not apply only to inflammations of the joints. We suppose that no other word in the patient's repertoire means arthritis. "Arthritis", in the counterfactual situation, differs both in dictionary definition and in extension from "arthritis" as we use it…. However we describe the patient's attitudes in the counterfactual situation, it will not be with a term or phrase extensionally equivalent with "arthritis." So the patient's counterfactual-attitude contents differ from his actual ones. (1979, p600-601)

If we take extension to be a constituent of the meaning of a term, then we will concur with Burge here that the meaning of S*'s term 'arthritis' differs from the meaning of S's term 'arthritis' (in later works the content of S*'s thought is called 'tharthritis'). We arrive, then, at Burge's conclusion, that the differences in the extension of the terms "spell differences in [the agents'] mental states" (1979, p601). When the content of an agent's mental states involve wide-content concepts, that content is constituted not only by states internal to the agent (since we assume that S's and S*'s internal states are identical), but also by facts in the agent's environment. Since concepts like arthritis are "widely individuated", there will be elements of the meanings of these concepts that are beyond the reach of an agent's first-person perspective. This will limit the agent's self knowledge, and the extent to which it is limited will be investigated in detail later. But here we can begin to see how mental content externalism puts some of the content of an agent's thought outside of the scope of an agent's first person perspective.

III The Conflict

Now that we have laid out the views that seem to be in conflict, we can examine precisely where the tension is between them and whether the conflict can be avoided. The apparent problem, once again, is this: If some mental content is wide, then some of the content of an agent's belief may not be internal to that agent. It seems to follow that some of the factors relevant to the justification of that agent's belief may not be appropriately internal to the agent in the way that justification internalism claims that they are.

The Compatibility of Mentalism and Mental Content Externalism

There are several varieties of justification internalism, and Conee and Feldman (2001) divide them into two categories, mentalist and accessibilist. What makes a view mentalist is that it requires that the factors relevant to the justification of a belief be internal to the mind of the agent. As Conee and Feldman write, Mentalism is "the view that a person's beliefs are justified only by things that are internal to the person's mental life" (2001, p233). This view does not specify precisely *how* something's being internal to an agent's mental life does the work of justifying their beliefs; it specifies only that the factors relevant for the justification of an agent's belief are internal to that agent's mental states.

Conee and Feldman's Mentalism is committed to two theses; "S" and "M" .

> S- The justificatory status of a person's doxastic attitudes strongly supervenes on the person's occurrent and dispositional mental states, events, and conditions.
>
> M- If any two individuals are exactly alike mentally, then they are alike justificationally, e.g., the same beliefs are justified for them to the same extent. (2001, p234)

Conee and Feldman take Mentalism to reflect the distinction between externalism and internalism found in philosophy of mind and ethics as well as epistemology, and they seem to sense a conflict between Mentalism and mental content externalism:

> What internalism in epistemology and philosophy of mind have in common is that being in some condition which is of philosophical interest- being epistemically justified in certain attitudes, or having certain attitudes with certain contents- is settled by what goes on inside cognitive beings. (2001, p233)

The association that Conee and Feldman draw between epistemological internalism and internalism in philosophy of mind shows that Conee and Feldman intend their use of 'mental' in S and M to be interpreted as including only narrow mental content, the content shared by denizens of Earth and Twin Earth who have thoughts they would express as, "This is water". "[A theory of justification] is internalism if and only if contingent factors external to the mind cannot make an epistemic difference" (2001, p234). According to this interpretation, we would take two agents who are "exactly alike mentally" to have precisely the same psychological content and/or physical constitution, but possibly inhabit two significantly dissimilar environments. We would also take the "mental states, events, and conditions" on which justification supervenes to be the narrow content of mental states, physical events internal to the agent, and the agent's psychological conditions.

This is certainly what Conee and Feldman had in mind when they proposed Mentalism as a form of epistemological internalism, but it does not follow from this that Mentalism, as its

theses S and M are stated, actually conflicts with mental content externalism. And on a more liberal reading of these two theses, we might interpret them in such a way that they could easily incorporate mental content externalism. We would take "mental" to include *both* narrow and wide content, as a mental content externalism views mental content. We could then hold both S and M without being committed to the view that mental states are individuated by conditions *internal* to an agent.

If we admit that some mental content is wide, we might hold both that the justification of an agent's beliefs strongly supervenes on that agent's mental states, and that the content of those mental states is sometimes determined by the agent's environment. If we take the "mental" in "mental content externalism" to be referring to the entire content of an agent's mental states, then the factors relevant to the justification of an agent's beliefs could be found in both the wide and narrow content of that agent's mental states. If we read "mental" in this way, there would be no conflict between Mentalism and mental content externalism.

We may become concerned that on this interpretation, one in which we take mental content externalism to be compatible with Mentalism, there will be no difference remaining between Mentalism and epistemological externalism; mental content externalism claims that external factors may partially determine the content of mental states, and justification externalism holds that some of the factors relevant to the justification of a belief are external to the agent. Mentalism's status as an internalist theory of justification looks questionable if we allow some of the content of an agent's mental state to be external to the agent, but perhaps this is merely a way of bringing attention to a general problem for Mentalism.

It seems as though the reason we value that the factors relevant to the justification of an agent's beliefs are internal to the agent's first-person perspective is that those factors would

thereby be accessible to the agent. Mentalism never guaranteed that our reasons would be accessible to us, but merely that we stand in a certain sort of relation to those reasons; not one of access, but a certain sort of ownership or internal possession. These reasons are our reasons because they are our mental states. Of course, we admit that many of our own mental states are not available to us- unconscious desires, for instance. If we really are interested in the desiderata laid out at the beginning of the paper, and this is our motivation for adopting an internalist theory of justification, then it seems that Mentalism will not be a satisfactory theory of justification, whether or not we take into account mental content externalism. We would not necessarily be able to determine, from our first-person perspective, that we are rational in forming our beliefs, or feel *confident* that our beliefs are true. Much more would need to be said to argue that Mentalism really is an unsatisfactory internalist theory in and of itself, and though I think there is reason to be suspicious of it and intend to investigate this matter thoroughly in future work, this matter is slightly beyond the scope of this paper.

In any case, it not does seem that Mentalism *guarantees* that we will have access to the factors that justify our beliefs, since Conee and Feldman specifically distinguish Mentalism from other internalist theories that emphasize our *access* to these factors. To save our original desiderata, in particular our desire to feel *confident* that the beliefs we hold are justified and *know* that we are rational in holding the beliefs that we do, we will need to examine whether mental content externalism is compatible with a stronger form of internalism, what Conee and Feldman call "Accessibilism". In order to do this, we will need to determine the extent to which we have *access* to our own thoughts, if some of the content of our mental states is wide, that is, outside our first-person perspective.

IV Self Knowledge

Introducing Mental Content Externalism and its Problem for Self Knowledge

The concern that justification internalism is incompatible with mental content externalism is related to, and perhaps arises from, a similar concern in the literature regarding self-knowledge and wide mental content. Paul Boghossian puts the point this way:

> Intuitively, the difficulty [that arises for self-knowledge in the face of mental content externalism] seems clear: how could anyone be in a position to know his thoughts merely by observing them, if facts about their content are determined by their relational properties? (1989, p11)

In "Content and Self-Knowledge", Boghossian argues that if we accept mental content externalism we cannot "know our own minds" (1989, p5). We take ourselves to know the content of our minds directly, Boghossian claims, without inference from anything such as our behavior or our other beliefs. There are two possible grounds for our self-knowledge: we either know what we think on the basis of introspection, or on no basis whatsoever. But if the content of our mental states is determined by factors beyond our own psychological or physical states, factors such as our physical or social environment, Boghossian claims one of two things follows.

1) If we believe that a faculty such as introspection gives us access to the content of our minds, this faculty will not tell us the relation that we stand in to the external world, and so

cannot reveal to us the content of our thoughts since that content is determined by such a relation.

2) If we adopt the position that we know the content of our mental states, but on no basis whatsoever, we encounter difficulties describing how it is that we know our own thoughts. (1989, p5)

This clearly could develop into a problem for justification internalism, since the view is that the factors which play a role in justification are the relations between an agent's beliefs or mental states, relations to which the agent has first-personal access.

Mental Content Externalism and Introspection

How does mental content externalism undermine *introspection* as an authoritative source of knowledge of our thoughts? Let's consider again Burge's example of the agent, S, who holds either a belief about *arthritis* or a belief about *tharthritis*, depending on S's social environment. As we say in the quote above, although S would express his belief as "I have arthritis in the thigh," no matter which context S is in, if S is in the actual world his belief will have the content (1) *I have arthritis in the thigh*, and if he is in the counterfactual world it will have the content (2) *I have tharthritis in the thigh*. By stipulation, S is not in a position to know whether the content of his belief is (1) or (2); he cannot determine which of the relevant alternatives holds in his case. Boghossian puts the point in this way:

> …S has to be able to exclude the possibility that his thought involved the concept arthritis rather than the concept tharthritis, before he can be said to know what his thought is. But this means that he has to *reason* his way to a conclusion about his thought; and reason to it, moreover, from evidence about his external environment

> which, by assumption, he does not possess. How, then, can he know his thought at all? –much less know it directly?" (1989, p14, italics mine)

S would need to reason to the conclusion that his environment is such that the content of his thought involves *arthritis* rather than *tharthritis*. But he could not know the relevant facts about his environment "by mere introspection. It would seem to follow, therefore, that I could not know the contents of my thoughts purely observationally: I would have to infer what I think from facts about my environment" (1989, p12). By introspection alone S cannot know whether the content of this thought involves *arthritis* or *tharthritis*, and if he cannot make this discrimination Boghossian claims S cannot know what he thinks.[2]

Addressing the point of whether arthritis and tharthritis are relevant alternatives, Boghossian draws on an analogy. "Someone may not be aware that there is a lot of counterfeit money in his vicinity; but if there is, the hypothesis that the dime-looking object in his hand is counterfeit needs to be excluded before he can be said to know that it is a dime" (1989, p14). If S has been switching back and forth between worlds in which the concepts *arthritis* and *tharthritis* are expressed by the term 'arthritis', then he is in an analogous situation as the agent who is in the presence of counterfeit coins, and whether S knows about his situation or not does not alter what the relevant alternatives are. "Epistemic relevance is not a subjective concept" (1989, p14), and through introspection S cannot discriminate between these relevant alternatives, therefore he cannot know the content of his own thoughts via introspection. Mental content externalism seems to leave introspection as an inadequate mode of gaining self-knowledge.

[2] There seems to be something that S knows; he knows that he has some thought, Φ, which he would express as "I have arthritis in my thigh". This would not seem to satisfy Boghossian, since S would not know the content of Φ, but this smaller piece of knowledge will play a major role in justification later in the paper.

So much, claims Boghossian, for introspective access to the content of our thoughts. Boghossian takes up a debate directly with Burge (1988) when he expresses the difficulties for the view that we can know our thoughts directly, through no faculty or process at all. Burge claims that our self-knowledge is not a matter of taking our thoughts merely as objects of other mental states. "When one knows that one is thinking that *p*, one is not taking one's thought (or thinking) that *p* merely as an object. One is thinking that *p* in the very event of thinking knowledgeably that one is thinking it" (1988, p654). Thinking knowledgeably that one is thinking that *p*, in Burge's terms, is to have a second order mental state that judges one's self to be thinking that *p*. It is to have a though such as *I judge: I am thinking that p*. Burge's position is that we know our thoughts "to be what [they are] by thinking [them] while exercising second order, self-ascriptive powers" (1988, p656), in our second order thoughts. In this way we can know our own thoughts without appealing to an activity such as introspection in order to gain knowledge of them. He writes,

> …perceptual knowledge of physical objects does not presuppose that one has first checked to insure that the background enabling conditions are fulfilled. The same point applies to knowledge of one's own mental events, particularly knowledge of the sort that interested Descartes. Such knowledge consists in a reflexive judgment which involves thinking a first-order thought that the judgment is itself about. The reflexive judgment simply inherits the content of the first-order thought. (1988, p656)

Burge uses the example of thinking that writing requires concentration. This is a first order thought, and to know that I am thinking this thought I need merely to have a second order thought that judges that I am thinking that writing requires concentration. The content of the first order thought is inherited by the second order thought, and therefore the second order

thought- *I judge: I am thinking that writing requires concentration*- is self verifying. It is not possible that my judgment could be wrong, since the thought about which the judgment is made is contained within that thought itself. Whatever the content of the first order thought may be, it is contained in the second order thought that both subsumes the first order thought and takes the first order thought as its object.

So on Burge's picture, whenever I am having a thought I can always make a veridical judgment that I am having that thought, and thereby know what thought I am having. My knowledge of my thoughts consists in my second order judgments about those thoughts, and not in a faculty of introspection I can exercise to observe what thoughts I am having. My self-knowledge depends solely on my having certain sorts of thoughts, second order self-verifying thoughts, and not on a process by which I observe my thoughts.

Boghossian raises several objections to Burge's position. The first is that Burge's theory does not seem to cover our "standing mental states" (1989, p21). We might make judgments concerning our beliefs, desires, fears, etc., but these judgments do not seem to be self-verifying.

> For example, [the thoughts] -I judge: I believe that writing requires concentration- or -I judge: I desire that writing require concentration- are not self verifying. I need not actually believe that writing requires concentration in order to think the first thought, nor actually desire that it require concentration to think the second. (1989, p21)

Thought this is no doubt correct, if we change the *attitude* of the thought it will fit Burge's theory. If I have the second order thought --I judge: I *think* that I believe writing requires concentration-- this thought *does* conform to Burge's picture. Boghossian makes the same objection regarding occurrently fearing that something is the case;

> Self-regarding judgments about what I occurrently desire or fear, for example, are manifestly not self-verifying, in that I need not actually desire or fear any particular thing in order to judge that I do. Thus it may be that –I judge: I fear that writing

> requires concentration—without actually fearing that it does. The judgment is not self-verifying. (1989, p21)

Again, Boghossian is correct. I may judge that I fear writing requires concentration without actually having this fear. I may be mistaken about my fear. But I cannot judge that I *think* that I fear that writing requires concentration without thinking that I fear that writing requires concentration. Although Boghossian states clearly that we take ourselves to "know about our beliefs and desires in a direct and authoritative manner" (1989, p21), our authority on our beliefs and desires has been in doubt since at least Freud's theories of psychology and the acknowledgement of unconscious fears, desires, expectations, etc. The reason that I may be mistaken about my fear is that *fearing* may not be the *sort* of mental state to which I have authoritative, first-person access. As Burge points out, "much of our self knowledge is similar to the knowledge of other's mental events. It depends on observations of our own behavior… And there is much that we do not know, or even misconstrue, about our own minds" (1988, p649). Boghossian's therapist may be in as good a position, or a better position, than Boghossian himself to know the nature of his fears. This does not impugn the special character of some of our self-knowledge: that we can know it authoritatively from our first-person perspective.

Boghossian claims that Burge's picture, at best, only guarantees knowledge of our thoughts that are "absolutely coincident" (1989, p21) with our second-order thoughts about them. "In other words, the second-order judgment will be self-verifying only if it literally incorporates the very thought about which it is a judgment" (1989, p21). Though Boghossian takes this to be a criticism of Burge's view, it is in fact precisely what Burge had in mind. Burge explicitly states that "the special epistemic status of these cases depends on the judgments' being made *simultaneously* from and about one's first-person point of view. The point of

view and time of the judgment must be the same as that of the thought being judged to occur" (1988, p658).

Burge's account guarantees that we can know our thoughts as they occur, and what makes my thoughts knowable is that at any time they occur I can judge that I am thinking them. This does not guarantee my knowledge of my deepest fears and desires, but no account of self knowledge in the offing claims to provide me with authority on these matters. It also does not *guarantee* knowledge of thoughts that have just occurred but are not longer occurring. But this is not a problem for a theory of self knowledge so much as it is for a theory of memory and its accuracy.

Burge's account gives us knowledge of our occurrent thoughts, insofar as we can simultaneously judge ourselves to be having those thoughts. He saves for us some sort of self-knowledge, but cannot give us complete knowledge of our mental states from the first-person perspective; though we have knowledge of our thoughts to some extent, our first person perspective cannot give us access to the *wide* content of these thoughts. We cannot individuate between the content of our mental states and the relevant alternatives to that content from within our first-person perspective. Burge, however, considers this to be irrelevant to our claim to know our own thoughts. We must come up with an interpretation, then, of *what* we know, if we cannot know the wide content of our thoughts.

Let's go back, once again, to Burge's example of the agent, S, who thinks he has arthritis in his thigh. Burge claims that S knows that he thinks he has arthritis in this thigh- he judges himself to have this first-order thought in a second-order, self-verifying thought. But what is it that S knows when he knows that he thinks that he has arthritis? We said above (in discussing Boghossian) that S cannot know the (entire) content of this thought, since his

thought may be about *arthritis* or *tharthritis*, depending on S's social environment. But there is *something* that S knows no matter what social environment S is in, in virtue of S having the appropriate second-order thought.

Let us call S's thought Σ, and the content of S's thought, including both wide and narrow content, Φ.[3] Since Φ depends on the relation S bears to his environment, S cannot know Φ except through empirical investigation. Φ may involve *arthritis*, or Φ may involve *tharthritis*; the nature of Φ is unavailable to S as the thought experiment stands. No matter whether Φ involves *arthritis* or *tharthritis*, S will express his thought (the content of which is Φ) as "I have arthritis in my thigh". Call the manner in which S would express this thought "Ψ". *Ψ* is what S knows in virtue of being able to have the second-order thought --I judge that I think that I have arthritis in my thigh--. S knows how he would *express* this thought, the thought that has the content Φ, and he would express this thought as "I have arthritis in my thigh", no matter whether he is in the actual or counterfactual social environment.

On my picture, the knowledge of S's thought, Σ, that Burge's account guarantees for S is Ψ, how S would express Σ, and not the content, Φ, of Σ.[4]

In other words, I can know that I have some thought Σ, and how I would Ψ this thought, that is, I would express it as "I have arthritis in my thigh". The content of Σ is Φ, which is not wholly available to me. Boghossian would not be satisfied with this sort of self-knowledge; when he asks, "how could anyone be in a position to know his thoughts merely by observing them, if facts about their content are determined by their relational properties?"

[3] Many philosophers would likely identify the content of a thought with the thought. I would prefer not to be committed to this identity, although for the purposes of this paper I doubt that it will make a difference either way.

[4] I intend Ψ to include not only the precise way in which S has or will express Σ, but also its grammatical transformations (excluding the substitution of coreferential terms, in cases where the subject does not know that these terms are coreferential).

(Boghossian, 11), he is clearly concerned with how one could be in the position to know the *content*, Φ, of his thoughts, and not how one would be in the position to know the manner in which one would express that thought. Burge, however, does not seem to share Boghossian's concern, and focuses on an agent's knowledge of his thoughts, rather than specifically the *content* of his thoughts. We might treat this as a mere terminological difference, but I think this would be a mistake and that the interpretation I propose accounts better for the difference between Boghossian's and Burge's positions on self-knowledge.

Burge certainly would not claim that we have first-person authority concerning the nature of Φ, the contents of our thoughts in their entirety. He writes, "One clearly does not have first-person authority about whether one of one's thoughts is to be explicated or individuated in such and such a way" (1988, p662). But one need not, on Burge's account, know Φ in order to know what one thinks. "Thus, I can know that I have arthritis, and know I think I have arthritis, even though I do not have a proper criterion for what arthritis is" (1988, p662). In whatever situation S is in, he can think, *I have arthritis*. But what S's thought has in common in both the counterfactual and actual situations is Ψ; the way in which S will express his thought in either situation is "I have arthritis". The content Φ of S's thought in the counterfactual and actual cases is completely different. As Burge says in "Individualism and the Mental",

> In the counterfactual situation, the patient lacks some, probably all, of the attitudes commonly attributed with content-clauses containing "arthritis" in oblique [opaque] occurrences. He lacks the occurrent thoughts or beliefs that he has arthritis in the thigh, that he has had arthritis for years… We suppose that in [this] case we cannot correctly ascribe any content-clause containing an oblique [or, opaque] occurrence of the term "arthritis". (1979, p600).

I take these considerations to support the interpretation I propose, that on Burge's account we know not the *content,* Φ, of our thoughts, since this includes both wide and narrow

content in some cases, but how we would Ψ those thoughts. I know that I have a thought, Σ, and the relation that I bare to the content of that thought, Φ, is Ψ, knowing the manner in which I would express Σ. The question for us to address now is whether this limited self knowledge, knowledge of Ψ, is sufficient for us to know how our thoughts relate to each other in such a way that justification internalism remains viable.

V How Limited Self Knowledge Threatens Accessibilism

We arrived at the conclusion above, that given the restrictions on self knowledge imposed by mental content externalism, S can know that he has a thought, and although he may not know the content, Φ, of that thought precisely, he can be related to that content in such a way that he knows the manner in which he would express, Ψ, that thought. This may be a problem for a stronger internalist view than Mentalism, or one stronger than any view which merely holds that justification supervenes on an agent's mental states, without specifying how that supervenience affects the agent's first-person relation to factors that justify their beliefs. One might hold that an agent needs access to the content, Φ, of her thoughts, and if this is the case accessibilist theories would not be viable in light of the restrictions mental content externalism impose on self-knowledge.

The accessibilist theory that Conee and Feldman describe they label (unsurprisingly) "Accessibilism", the view that "the epistemic justification of a person's belief is determined by things to which the person has some special sort of access" (2001, p233). If the factors that justify our beliefs are only factors to which we have access via our first-personal relations to our mental states, then the restrictions that mental content externalism put on our relation to the content of our beliefs might undermine an accessibilist view. Perhaps we must know the entirety of the content, Φ, of our mental state to see what relation that mental state bares to another. The restriction of an agent's knowledge to only the manner in which he

would express his mental state, Ψ, instead of encompassing the entire content, Φ, on this interpretation would keep him from having access to the relations between his mental states. This, of course, would put the factors relevant to justification *beyond* the scope of the agent's first person perspective, and thereby undermine the general accessibilist position, that the factors relevant to the justification of an agent's belief are factors to which an agent has *access*.

I propose that although mental content externalism seems to threaten Accessibilism, this is only a *prima facie* threat. My suggestion is that knowing how one would Ψ ones thought Σ, even though one does not have access to the content Φ of Σ, will give an agent knowledge of the justificatory relations between Σ and other of his thoughts. Granted, knowledge of this relation is not knowledge of the entire way in which one's thoughts are related- an omniscient being would have more thorough knowledge of the relations between his thoughts than our agent S does. My proposal is that S's knowledge of the Ψ of his thought Σ provides S with access to the relations between his thought Σ and his other thoughts ($\Sigma_1 \ldots \Sigma_n$) such that he will know whether Σ is justified by $\Sigma_1 \ldots \Sigma_n$. The relations between the Ψ's of Σ's, not the Φ's of Σ's, are the factors relevant to the justification of Σ. Even under the restrictions that mental content externalism could put on self knowledge, I believe that Accessibilism will still satisfy the desiderata laid out at the beginning of this paper.

How Real is the Threat?

Bonjour senses an incompatibility between mental content externalism and justification internalism when he writes,

> "…if part or all of the content of a belief is inaccessible to the believer, then both the justifying status of other beliefs in relation to that content and the status of that content as justifying further beliefs will be similarly inaccessible, thus contravening the internalist requirement for justification." (1992, p136)

James Chase (2001) addresses the worry that Bonjour expresses here and defends the compatibility of mental content externalism and justification internalism. He holds that there is an equivocation at play in the use of "inacessible" in the statement above. We might view BonJour's claim in a similar way. On our account of the limitation that mental content externalism puts on self-knowledge, we would treat BonJour's statement as not distinguishing between the agent's knowledge of the Ψ of that belief and the content, Φ, of that belief. In cases in which a belief involves a wide content concept, an agent will not know the Φ of that belief, and therefore does not have the same thorough knowledge of the relations between his mental states that an omniscient agent would have. But the agent does have access to the Ψ of each of his beliefs, and to the relations between the Ψ's of his beliefs.

Chase precisifies Bonjour's claim in the following way;

> **B1)** If Content Externalism is true then there can be an agent A with belief B such that part or all of the content of B is not internally available to A.
> **B2)** If an agent A with belief B is such that part or all of the content of B is not internally available to A, then the justification relations B stands in with other beliefs of A's are not internally available to A.
> **B3)** If agent A with belief B is such that the justification relations B stands in with other beliefs of A are not internally available to A, then not all factors relevant to the justificaiton of beliefs of A are internally available to A.
> **C'')** If Content Externalism is true then Justification Internalism is false. (2001, p237)

The equivocation, as Chase spells out Bonjour's argument, is in the notion of internal availability. In B1, clearly, what is not internally available to agent A is part (or all) of the content, Φ, of belief B. In B2, it is the justification relation(s) between one belief and another which are not internally available to the agent. However, from the fact that part of the content, Φ, of A's belief is not available to A, it does not follow that the the *justification relations* B stands in to other beliefs are entirely unavailable to A**.** On my proposed account, the relations between the Ψ's of Σ's, not the Φ's of Σ's, are the factors that serve to justify Σ. So long as the Ψ's of A's beliefs are available to A, the justification relations in which B stands to other beliefs will be available to A as well.

It would be an error to assume that the internalist would take the relations between the Φ's of Σ's as the justificatory relations between Σ's. I propose that internalists have *always* been committed to the view that only the relations between the Ψ's of beliefs are relevant to the justification of a belief; although they have not thus far explicitly distinguished between content Φ and the corresponding Ψ of the thought with content Φ.

Consider the case of S and S*. In taking S and S* to be equally justified in their beliefs, internalists are treating the relations between the Ψ's of Σ's as the justificatory relation between the Σ's of S and S*: it is the Ψ of Σ that S and S* share in common. To put the point more strongly, and I think accurately, the Ψ's of S's and S*'s Σ's are *identical,* and thus the justificatory status of S's and S*'s Σ's are *identical* as well. Internalists hold that the envatted and unenvatted agents, with identical internal/psychological histories, to be justified to the same extent and in the same way in their beliefs, Σ, if those beliefs have the same Ψ.

We need not treat S and S* as holding the same belief; clearly, given the difference in the content (Φ and Φ*) of Σ and Σ*, there is at least *some* difference between their beliefs. But

on this view it does not matter whether Σ and Σ* are the same belief of different beliefs. We have distinguished between the Φ of a thought and the Ψ of that thought, and thus we can identify what it is that Σ and Σ* share in common- their Ψ's.

Although taking into account the restrictions that mental content externalism places on self-knowledge requires that the internalist rearticulate the factors that are relevant to the justification of a belief in order disambiguate between Φ and Ψ, this rearticulation is only the clarification of an ambiguity; I believe it is not a significant modification of the internalist position.

Chase's response to the problem that BonJour identifies employs a Brain-in-a-Vat thought experiment. This response to Bonjour, as one would expect, compares the response a justification internalist would give to the problem of mental content externalim to other cases in which external factors relevant to an agent's belief are unavailable to that agent. According to the internalist position, the justification of the envatted agent's beliefs depends neither on how those beliefs match up to the world (their truth or falsity) nor on the wide content of the agent's beliefs. The elements of an agent's mental state that are relevant to the justification of an agent's beliefs are those that are shared between the envatted and unenvatted agent. Though I believe Chase's response to be a successful reply to BonJour's proposed problem, it does not articulate *what* element of an agent's belief is held constant between contexts in which he is a brain in a vat and when he is in the ordinary world, except to state that it is the element of the agent's thought to which the agent has access. The problem with Chase's answer to Bonjour is not that it is inaccurate, but that it is vague.

Burge's example of the patient who has beliefs about arthritis and to what ailments the term "arthritis" applies is more unique to the issue of mental content externalism. We

stipulated that S and S* are identical in their histories, physical constitution, dispositions to behave, etc, and differ only in their social environments. And we drew the conclusion that S's belief expressed as "I have arthritis in the thigh" and S*'s belief also expressed as "I have arthritis in the thigh" have different content, Φ, although there is nothing within the agents themselves (the internal properties of the agents) that accounts for this difference in Φ. The Φ of S's belief involves *arthritis*, while the Φ of S*'s belief cannot involve *arthritis* since there is no term that means *arthritis* in his linguistic community. Instead, the content Φ of S*'s belief involves *tharthritis*.

How does this affect the justificatory status of S's and S*'s beliefs? As internalists, to determine if either of these beliefs are justified we would look to the relations that these beliefs bare to additional beliefs or other justifying factors that are internal to the agent. As we set up the example above, S came about his belief through causal reading, conversation, etc., in the way that most lay people arrive at their beliefs about medical conditions. S believes that the materials he has read are good sources for gathering information on medical conditions, he believes that the people with whom he has spoken about arthritis are intelligent and fairly well informed, he believes that he has read these materials and heard these people correctly, etc. It is the similarity in the histories listed above that accounts for the fact that S and S* would both Ψ their beliefs (with different Φ's) as "I have arthritis in my thigh". That S and S* would Ψ their beliefs in the same way is not coincidental: the Ψ derrives from S's and S*'s identical histories. Furthermore, were the histories of S and S* not identical, we would not be concerned with whether their Σ's were both justified in the same way and to the same extent.

How do S and S* come to know the way in which their beliefs Σ and Σ* are related to their other beliefs? S and S* will have certain additional beliefs, which were grounded in related experiences in the world (readings, discussions with intelligent friends, etc.) which would each be Ψ'd in its own way. The Ψ of each belief will be related to the Ψ's of other beliefs, as the Φ of a belief would be related to the Φ of another belief. And the Ψ of a belief, Σ, will be related to the Φ of Σ such that the Ψ of Σ will not vary unless the Φ of Σ varies as well. S and S* both understand something about *arthritis* and *tharthritis*, or we would have no reason to attribute to them thoughts with Φ's that involve *arthritis* and *tharthritis* respectively. As Burge notes, "It is a truism that to think one's thoughts, and thus to think cogito-like thoughts, one must understand what one is thinking well enough to think it" (1979, p662). S and S* will both have *some* understanding of Φ. They will share their understanding of the Φ of their thoughts in common, which explains why they will Ψ their Σ's identically, as "I have arthritis in my thigh". These Ψ's do not arise in the speaker at random, but rather from the histories of S and S*, histories that S and S* share in common, barring the subtle differences in their environments. These histories both explain the acquisition of their thoughts Σ and Σ* with contents Φ and Φ*, and why S and S* will Ψ Σ and Σ* identically.

So the Ψ of an agent's thought will correspond to the content Φ of that thought in a systematic way produced by the histories that agents like S and S* share in common. Other of the agent's thoughts with their unique Ψ's that also correspond to their contents Φ in systematic ways will be related to each other not merely by their contents Φ (which are relations to which an omniscient agent would have access) but their Ψ's as well. The agent will have access to the Ψ's of his thoughts, and will thereby have access to these relations

between the Ψ's of his thoughts. It is the Ψ's of thoughts that agents in thought experiments such as Putnam's Twin Earth share in common across worlds, and it is the Ψ's of thoughts that are held constant between situations in which an agent is envatted and when that agent is in the actual world. Internalists have always maintained that it is what the agents' mental states share in common in these thought experiments that is relevant to the justification of these agents' beliefs. The properties that otherwise identical envatted and unenvatted agents share are the properties that factor into the justification of an agent's beliefs. Now we see both what that the agent's mental states share in common, their Ψ's, and how these Ψ's can be related to each other in such a way that the agent has sufficient knowledge of the relations between his beliefs to justify those beliefs.

Works Cited

BonJour, Lawrence. 1992. 'Externalism/Internalism'. In J. Dancy, E. Sosa. Editors. *A Companion to Epistemology.* Oxford. Blackwell. pp. 132-136.

BonJour, Lawrence, 2002. *Epistemology, Classic Problems and Contemporary Responses.* Oxford. Rowman and Littlefield.

Boghossian, Paul. Spring 1989. 'Content and Self-Knowledge'. *Philosophical Topics*, Vol XVII. No. 1. pp. 5-26.

Burge, Tyler. 1979. 'Individualism and the Mental'. P. French, T. Uehling, and H Winston. Editors. *Studies in Metaphysics*. University of Minesota Press. In Chalmers, David. 2002. *Philosophy of Mind, Classical and Contemporary Readings*. Oxford. Oxford University Press.

Burge, Tyler. Nov., 1988. 'Individualism and Self-Knowledge'. *The Journal of Philosophy*. Vol 85. No. 11. Eighty-Fifth Annual Meeting American Philosophical Association, Eastern Division. pp. 649-663.

Chase, James. June 2001. 'Is Externalism about Content Inconsistent with Internalism about Justification?'. *Australasian Journal of Philosophy*. Vol. 79. No. 2. pp. 227-246.

Conee, Earl and Feldman, Richard. 2001. 'Internalism Defended'. in Kornblith, Hilary. Editor. *Epistemology: Internalism and Externalism*. Cambridge. MIT Press.

Putnam, Hillary. 1975. 'The Meaning of Meaning'. In K. Gunderson. Editor. *Language, Mind, and Knowledge.* University of Minesota Press. pp. 131-193. In Chalmers, David. 2002. *Philosophy of Mind, classical and contemporary Readings*. Oxford. Oxford University Press.