CSLI LECTURE NOTES
NUMBER 85

Externalism and Self-Knowledge

edited by
Peter Ludlow &
Norah Martin

CSLI Publications
Center for the Study of
Language and Information
Stanford, California

CSLI Publications
Center for the Study of Language and Information
Leland Stanford Junior University
02 01 00 99 98 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Externalism and self-knowledge / edited by Peter Ludlow & Norah Martin.
p. cm.–(CSLI lecture notes ; no. 85)
Includes bibliographical references and index.
ISBN 1-57586-107-0 (hardcover : alk. paper).—ISBN 1-57586-106-2
(pbk. : alk. paper)
1. Externalism (Philosophy of mind) 2. Self-knowledge, Theory of.
I. Ludlow, Peter, 1957– . II. Martin, Norah, 1962– . III. Series.
BD418.3.E87 1998
128' . 2–dc21 98-14147
CIP

∞ The acid-free paper used in this book meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

The painting on the cover of the paperback edition of this book, Edward Hopper's *People in the Sun*, appears courtesy of the National Museum of American Art, Smithsonian Institution. Gift of S.C. Johnson & Son, Inc.

Contents

20

Self-Knowledge and Closure

Sven Bernecker

If 'privileged self-knowledge' means knowledge that is not based on investigations of one's environment, we certainly seem to possess privileged self-knowledge of some of our conscious and occurrent thoughts. Nevertheless, some have questioned our ability to authoritatively know our mental states. One reason for doing so derives from externalism (or anti-individualism), the view that the content of an intentional state is fixed in part by the environment external to the believer. Because mental content supervenes on extrinsic relations it is called 'broad' or 'wide' content. The leading argument for the incompatibility of privileged self-knowledge and externalism claims that privileged self-knowledge is incompatible with the kind of twinning and switching scenarios on which externalism is based. It is argued that this argument is flawed and that privileged self-knowledge is consistent with externalism.

The advocates of the incompatibilist argument discussed in this paper are, among others, Boghossian, BonJour, Brueckner, and McGinn.[1] The argument proceeds from a slow switching thought experiment introduced by Burge (1988, pp. 652ff.) where an agent, S, is switched from Earth to Twin Earth and remains there for some time. The only difference between the two planets is that Twin Earth doesn't have any water. Instead of water there is a liquid that superficially resembles water but which has the chemical formular XYZ. Practically everyone agrees that after a while (how long is unclear), tokens of 'water' in S's mentalese will

[1] Cf. Boghossian (1989, pp. 13–4; 1992, pp. 17–21; 1994, pp. 36–40), BonJour (1991, p. 339), Brueckner (1990, pp. 449–51; 1992, pp. 206–7; 1994), and McGinn (1989, pp. 82–94). For a discussion of other incompatibilist arguments see my 1996a.

cease to mean water (H_2O) but will instead refer to XYZ. If we want to express S's word 'water' in English, we have to coin a word – 'twater', say. The content of the thought S expressed on Earth by saying "I am thinking that water is wet" is different from the thought content he expresses by the same sentence uttered on Twin Earth. However, since nothing internal to S distinguishes the two worlds, he cannot know by introspection which one he is in. But then, doesn't it follow that he lacks privileged knowledge of the content of his thought? Incompatibilists say 'yes' for they maintain that "S has to be able to exclude the possibility that his thought involve[s] the concept [water] rather than the concept [twater], before he can be said to know what his thought is" (Boghossian 1989, p. 14). Since introspection doesn't tell him whether he is entertaining a water thought or a twater thought, he lacks privileged self-knowledge of (some of his) thought contents. And since Twin Earth scenarios are the primary motivation for externalism, it follows that externalism is inconsistent with the doctrine of privileged self-knowledge. This argument can be parsed into the following steps:

(1) S is switched from Earth to Twin Earth. In situations where S used to think water thoughts he now entertains twater thoughts.

(2) Since S is supposed to remain physically identical through the switching, he cannot introspectively discriminate water thoughts from twater thoughts.

(3) Water thoughts and twater thoughts differ in broad content.

(C) Thus, if S has privileged self-knowledge, what he has knowledge of is not broad content.

It is important to realize that the force of this reasoning depends neither on S's being ignorant of the existence of Twin Earth nor on his forgetfulness. Given externalism, it is possible for S to know his thoughts at time t_1 (before the switch), forget nothing, be informed about the switch, and yet at some later time t_2 be unable to know *introspectively* what the contents of his thoughts were at t_1. Boghossian (1989, p. 23) argues "the only explanation [...] for why S will not know tomorrow what he is said to know today, is [...] that he never knew". To know what he thought at t_1, S would have to discover what environment he was in at that time and how long he had been there. But these things cannot be known authoritatively.

Incompatibilists conclude that externalism cannot account for privileged self-knowledge.

As it stands, the incompatibilist argument is invalid. What is missing is an explanation for why being able to distinguish non-empirically water thoughts from twater thoughts should be a necessary condition of privileged self-knowledge. To complete the argument the following premise has to be added:

(4) Privileged self-knowledge is closed under known entailment.[2]

It is the *closure principle* applied to privileged self-knowledge that does all of the work in the above incompatibilist argument. The principle of closure under known implications states that if S knows P, and S knows that P entails Q, then S knows Q. This principle is commonly used to back skepticism regarding our ability to know about the external world: If I know that this is water, I also have to know I am not on Twin Earth. But I don't know that I am not on Twin Earth, since if I were, things would seem exactly as they seem now and this would not be water. It follows that I don't know that this is water. Whereas the Cartesian skeptic employs epistemic closure to undermine knowledge of the external world, Brueckner uses the closure principle to show that, given externalism, privileged self-knowledge is impossible: "[M]y knowing that I am thinking that some water is dripping requires that I know that I am not thinking that some twater is dripping," for self-knowledge implies the ability to rule out counterfactual thought contents. But I don't know authoritatively that I am not thinking that some twater is dripping, for, "if I were a brain in a vat [in a waterless world], things would seem exactly as they now seem (and have seemed)" (Brueckner 1990, pp. 450, 448; cf. 1992, pp. 206–7). Therefore, Brueckner and other incompatibilists conclude, I don't know authoritatively that I am thinking that some water is dripping, i.e., I don't know the content of my thought.

Regarding knowledge of the outside world most epistemologists agree that the closure principle is too strict to be convincing. If knowing that P would require the elimination of *every* alternative to P, as suggested by the closure principle, we could never know anything about the world around

---

[2] Instead of premise (4) some incompatibilists assume the following premise (4'): Privileged self-knowledge of thought contents requires that one can determine non-empirically sameness and difference in thought contents. Since premise (4') is a derivative form of the closure principle it will not be discussed separately. For a good discussion of premise (4') see Owens 1992.

us.[3] A much more plausible view is that knowledge requires the elimination of *relevant* alternatives only. This position is commonly called the '*relevant alternative account of knowledge*'. Which alternatives are relevant? Relevance has to do with the kind of possibilities that actually exist in the objective situation.[4] In an ordinary case of claiming to know that some animals in the zoo are zebras, to use Dretske's example, the alternative that they are cleverly painted mules is not relevant. Thus, one can truthfully claim to know that they are zebras despite one's inability to rule out this fanciful alternative. But in some extraordinary cases, the painted mules hypothesis is a relevant alternative; and then we have to eliminate this alternative to know that what we are seeing are really zebras.

Warfield (1992) has offered an objection to the incompatibilist argument by applying the relevant alternative approach to self-knowledge. Why, he asks, should the conditions of knowledge of one's own thoughts be any stricter than the conditions on knowledge of the external world? If knowledge of the external world isn't closed under known implications, why should privileged self-knowledge be closed? Warfield argues that the incompatibilist reasoning, by itself, fails to rob us of privileged self-knowledge. Only if switching cases were relevant, we would fail to know what it is we are thinking. In ordinary circumstances, however, the mere possibility of switching cases no more undermines my self-knowledge than the

---

[3] This is not the only reason to get rid of the closure principle: Knowledge is closed under known logical implications only if each necessary condition for knowing is so closed. Any correct analysis of knowledge must contain a belief condition and a truth condition. Truth is obviously closed under known implications and belief is arguably too. But what about the justification condition? Assuming that there is a justification condition for knowledge, many epistemologists have maintained some kind of closure principle for justification (cf. Hooker 1973). However, given reliabilism, the proper analysis of knowledge doesn't yield a justification condition. The justification condition is replaced by what might be called the 'reliability condition': a belief that P qualifies as knowledge only if it is a reliable indicator of the truth that P. Advocates of closure would have to show that the reliability condition is closed under known implications; and this seems hard to do.

[4] On the notion of relevant alternatives see Dretske (1981) and Goldman (1976). Apart from the objective reading of relevance there is a subjective reading according to which the attribution of knowledge is relative to the conversational context. For example, the sentence "S knows that he is sitting at his desk" can be true or false depending on whether the claim is taken to imply that S is sitting at his desk (not yours), or that he is sitting at his desk (not standing), or that he is sitting at his desk (not his armchair). The subjective notion of relevant alternative is the basis of contextualism.

mere possibility of me possessing counterfeit coins undermines my knowledge that I have a dime in my pocket.

In response to Warfield, Ludlow (1995a) has argued for the prevalence of switching scenarios. While Twin Earth scenarios are normally about tokens of the same word type having different meanings in different worlds, Ludlow examines the reverse case – tokens of distinct word types having the same meaning in different worlds. And instead of moves between Earth and Twin Earth, Ludlow considers moves between language communities, social groups and institutions: Biff who knows very little about vegetables moves between the United Kingdom and the United States. In the United Kingdom, Biff expresses a particular thought by saying "Chicory is tasty". When the same internal episode takes place in the United States he says "Arugula is tasty". Biff is unaware that the contents of both thoughts are identical. Moreover, this fact isn't detectable purely on the basis of introspection. According to Ludlow, Biff is a victim of a slow switching case. Since we routinely move between social groups and institutions and because Biff-like cases can be applied not only to a wide variety of nouns (e.g., 'chips' and 'crisps', 'gasoline' and 'petrol') but to any part of speech (e.g., 'latino' and 'hispanic'), Ludlow concludes that, contrary to Warfield's contention, slow switching cases are indeed prevalent.

I applaud Warfield's criticism of incompatibilism. Just as in ordinary cases I can know that some animals in the zoo are zebras without having to know that they are not cleverly painted mules, I can normally know that I am having a water thought without being able to rule out the possibility that I am having a twater thought. Since there is no reason to suppose that the requirements for knowledge of one's thought contents are any stricter than those for knowledge of the external world, I claim that, in ordinary cases, self-knowledge isn't closed under known implications. While agreeing with the relevant-alternatives account of knowledge, I think it is unlikely that incompatibilism can be refuted in this way. The dispute on the prevalence of slow switching cases easily degenerates into an idle exchange of burden of proof arguments. Fortunately however, we do not have to enter into the discussion on whether twater thoughts are relevant or irrelevant alternatives, for they aren't (entertainable) alternatives at all. Let me explain.

Burge (1988, 1996) and others have provided a convincing case for the compatibility of externalism and privileged self-knowledge. To know that I am thinking that, say, water is wet, I don't need to first acquire knowl-

edge of either how experts in my community use 'water' (social externalism) or of the kind of substance I was in contact with when I learned 'water' (causal-essentialist externalism). The reason is that the content of the first-order thought ("water is wet") is automatically contained in the content of the second-order thought ("I believe that water is wet") and the contents of both thoughts are determined by the same causal relations of which one may be ignorant. So no matter which planet one resides on, as long as the first-order thought and the second-order thought are entertained simultaneously, the content of the second-order thought cannot come apart from the first-order thought by which it is causally sustained. I refer to this compatibilist line as the '*inclusion theory of privileged self-knowledge*'.

The inclusion theory of self-knowledge relies on a reliabilist conception of knowledge. *Reliabilism* is the view that to know something all that is required is that the belief-fact link is reliable. The subject doesn't need to know that it is reliable. Reliabilism replaces the traditional notion of justification by a nomic (e.g. causal or counterfactual) or other such external relation between belief and truth. Therefore, to have self-knowledge a subject need only, as a matter of fact, stand in some causal relation to his first-order states; he need not know that he does. Reliabilism stands opposed to epistemic internalism, the view that knowledge requires cognitive access to the justificatory procedure and evidence on which it is based. Some versions of internalism are committed to the KK-principle, i.e., the view that knowing entails knowing that one knows. Without the assumption of reliabilism, the inclusion theory of self-knowledge might explain why second-order thoughts of the form "I am believing that water is wet" are necessarily true, but it would fall short of accounting for introspective *knowledge*.[5]

The combination of reliabilism and the relevant alternative account of knowledge yields a powerful objection to the incompatibilist argument under consideration.[6] Given the inclusion theory of self-knowledge, the

---

[5] Burge endorses reliabilism regarding introspective knowledge, but he doesn't assume it as a general doctrine of knowledge.

[6] Although most reliabilists deny the closure principle, reliabilism and closure are, in principle, compatible. Suppose a version of reliabilism that rests on the following counterfactual condition: If P were false, then S would not mistakenly believe that P. Now suppose S's claim to know logically implies the proposition that S is not a brain in a vat (BIV). If S were a BIV, then, presumably, S would mistakenly believe that he is not a BIV. Hence, he doesn't know that P. Thus, if the belief-forming process counts as reliable iff it actually yields a suf-

hypothesis according to which I am having twater thoughts while thinking that I have water thoughts simply isn't a (entertainable) possibility. When I am on Earth thinking earthian concepts, I *cannot* believe that I am thinking that twater is wet for I don't have the concept of twater available; so this concept cannot figure in any of my mental states. Analogously, when I am on Twin Earth, I *cannot* mistakenly believe that I am entertaining water thoughts. No matter how often I am switched between Earth and Twin Earth, I will never erroneously think that I am having water thoughts while in fact I am having twater thoughts and vice versa. Privileged self-knowledge is therefore immune to skeptical arguments from switching and twinning scenarios.

This is the central compatibilist argument. In the remainder of the paper I will consider three incompatibilist rejoinders, none of which is convincing.

(1) Apparently the compatibilist argument above presupposes that the switching or twinning of an agent S brings about a *complete change* of his concepts and thought contents. For it is claimed that S cannot mistake his twater thoughts for water thoughts and vice versa because on Twin Earth he doesn't have available the concept water and on Earth he doesn't have available the concept twater.

Now Boghossian (1992, pp. 19–21; 1994, pp. 38–9) has challenged the idea that switching and twinning results in a complete change of the agent's concepts and contents. The suggestion is that when on Twin Earth S remembers water thoughts from the time he was living on Earth, then, despite being tokened on Twin Earth, these tokens of 'water' occurring in memories retain their earthly interpretation. Thus it is not impossible for S to have available both water thoughts and twater thoughts at the same time. But then switching and twinning hypotheses may indeed represent relevant alternatives to one's privileged self-knowledge. It might happen that what S takes to be a twater thought is in fact a remembered water thought. And since twater thoughts and water thoughts cannot be

---

ficiently high ratio of true beliefs both in the actual and the counterfactual situation, then reliabilism is compatible with closure. Moreover, although most advocates of the relevant alternative account of knowledge endorse reliabilism, the relevant alternative account is, in principle, compatible with epistemological internalism. It is possible to hold a sort of internalist theory of knowledge that highlights the role of evidential justification (and reject reliabilism) and then go on to deny closure via embracing the notion of a relevant alternative. I owe this point to Tony Brueckner.

discriminated on the basis of introspection alone, S would lack privileged self-knowledge of what he is thinking.

As Ludlow (1995b) has pointed out, this defense of incompatibilism rests on a conception of *memory* which conflicts with the spirit of externalism. Any externalist conception of memory has it that the content of a memory is in part dependent on systematic external relations. But which relations – past or current ones? Boghossian assumes that S's content of his memory fixed at time t_1 remains frozen up to some later moment of recollection. But if externalism is right and mental contents are determined by external facts, why not suppose that the content of a memory is fixed at the time recollection takes place, since it is the embedding circumstances of that memory which are crucial to the fixing of its content? On this view of memory, S's memory of event W at t_1 is different from his memory of W at t_2, if the content-determining environment changes. Just as slow switching can bring about changes in mental content, it can bring about changes in memory content. It is therefore impossible that, while residing on Twin Earth, S's tokens of 'water' occurring in memories can retain their earthly meaning.

Ludlow's externalist notion of memory has met criticism. Burge (1996, p. 97n), for example, has warned against "overrat[ing] the extent to which the content retrieved in memory is sensitive to *immediate* environmental context".[7] The problem some philosophers are having with Ludlow's concept of memory is this: At t_1, while living on Earth, S thinks that water is wet. At t_2, after having been transported to Twin Earth, S thinks "At t_1 I thought that water is wet". How can the thought at t_2 involve the concept twater and, at the same time, express a memory of an earthly thought? For a state to be a memory doesn't its content have to involve concepts that apply to the remembered event? Ludlow's concept of memory is said to be implausible for it suggests that, in an important sense, it is beyond one's control whether or not one can remember one's previous thoughts. Moreover, in conjunction with the prevalence of slow-switch-

[7] My emphasis. Cf. Brueckner (1997, pp 5–6.), Hofmann (Mscr.), and Ludlow (1996). Burge's (ch. 21 in this volume) most recent thoughts on preservative memory (which I read after having finished this chapter) support my idea that memory doesn't require the ability to distinguish original thoughts from new twin thoughts.

ing scenarios, Ludlow's notion of memory entails that one can seldom remember one's earlier thoughts.

The problem with this objection to Ludlow's notion of memory is the assumption that a memory state *necessarily* contains the content and concepts of the relevant earlier state. Why should we suppose that our memory is sensitive to twinning scenarios given that water thoughts and twater thoughts are phenomenologically, functionally and introspectively indistinguishable? What would be the evolutionary utility of designing memory in such a way as to make it respond to differences that, introspectively speaking, don't make a difference? What I am suggesting then is that the job of memory, rather than to replay previously recorded contents, is to provide information about past states relative to the present environmental conditions. The transfer of contents and concepts across time might be a sufficient condition for memory but it falls short of being a necessary condition.

Even if there was a strict proof that the content of a memory is fixed at the time recollection takes place, there might still be situations in which an agent possesses both concepts, water and twater, at the same time. Imagine, for example, that while being transported from Earth to Twin Earth S is very thirsty. He continuously thinks "I have been wanting a drink of water for the last two hours". Now it could be argued that once S has reached Twin Earth (and has stayed there for a while) S's term 'water' in "I have been wanting a drink of water for the last two hours" is ambiguous in that it refers to both H_2O and XYZ. But if S can have both concepts – water and twater – at the same time, isn't it then possible that he mistakes his twater thought for a water thought and therefore lacks privileged self-knowledge of his thought contents? And if this is so, doesn't it follow that privileged self-knowledge is inconsistent with switching and twinning scenarios?

For this objection to be convincing the compatibilist argument presented above would have to presuppose that the switching and twinning of an agent S brings about a *complete change* of his concepts and contents. But compatibilism isn't committed to this presupposition. Even if S had available water *and* twater concepts simultaneously, he still couldn't erroneously think that he is having a water thought while in fact he is having a twater thought and vice versa. For given the inclusion theory, if the first-order thought isn't just an object of reference but is part of the higher-order cognition itself, the concepts of the first-order thought cannot come apart from those of the second-order thought. So when S is think-

ing a water thought, his second-order thought involves the concept water; and when he is thinking a twater thought, his second-order thought involves the concept twater. So no matter what he does, he cannot get his first-order thought wrong.[8]

(2) The inclusion theory of self-knowledge shows no more than that the hypothesis according to which I am having twater thoughts while thinking that I am having water thoughts is not an *entertainable* possibility. In other words, the scenario envisioned by the incompatibilist involves a *pragmatic* contradiction. But this doesn't mean that the scenario could not be true. This would only follow if it contained a *logical* contradiction. The hypothesis that I have alternative contents is indeed possible, although not thinkable. Thus, we are even worse off than we thought. We lack the necessary concept to express what may very well be true, and our causal circumstances make it impossible for us to acquire them.

The incompatibilist could use this reasoning to construct an argument to the effect that our ordinary claims to introspective knowledge are unjustified. He could suggest that the class of relevant alternatives is not restricted to actual thought contents but also includes possible thought contents. The idea is that the unthinkable and thus purely *abstract* possibility of alternative contents is enough to destroy privileged self-knowledge. Even if one cannot coherently think that one mistakes one's present water thoughts for twater thoughts, if one dwells on switching cases long enough, the level of scrutiny will rise and one will find oneself unwilling to claim to know that one is occupying a water thought (as opposed to a twater thought). Self-knowledge can be undermined just by knowing in the abstract that there are alternative contents, without ever knowing what those alternatives are and without ever being able to verify their existence via introspection. The situation is similar to that of a person who has never seen a counterfeit coin and doesn't know what a counterfeit looks like but who, after having read Descartes' *Meditations on First Philosophy*, worries that the dime-like looking object in his pocket is a counterfeit. This person might find himself unwilling to claim to know that he has a dime in his pocket.

To see what is wrong with this incompatibilist rejoinder one needs to realize that there are two very different phenomena that can be thought of as the contextual relativity of knowledge. One characteristic of knowledge is that it is determined relative to the *extra-evidential context* of the sub-

[8] I owe this point to John Perry and Ken Taylor.

ject. Subjects living in different environments can possess the same evidence for the truth of a certain proposition P, and one of them knows that P while the other one fails to know it. It is this notion of contextual relativity of knowledge that underlies (my objectivist reading of) the relevant-alternative account of knowledge. An alternative construal of contextual relativity says that knowledge *attributions* are dependent on the *conversational context*. On this view, the attributor's purposes, intentions, and presuppositions of the epistemic subjects play a role in setting the standards of relevance. I could say of someone in one conversational context that he knows that P and could then go to a different conversational context and say of him that he doesn't know that P. The same is said to apply to cases of self-attribution of knowledge. While reliabilism emphasizes the extra-evidential context of knowledge, contextualism stresses the conversational context.[9]

The above incompatibilist rejoinder rests on a confusion of conditions for knowledge attributions with conditions for knowledge. Even if slow switching scenarios destroy one's attributive self-knowledge, they don't thereby destroy the introspective knowledge one has of one's mental condition. For just because one cannot justifiably *say* that one knows what one is thinking doesn't show that in fact one doesn't know what one is thinking. Having reasons for knowing and having reasons for self-attributing knowledge are two quite different things. One might not have reasons for attributing knowledge to oneself and still know. In other words, one can know that one is thinking that P and not be able to justifiably say that one knows what one is thinking. And therefore, the abstract possibility of alternative thought contents is unable to undermine privileged self-knowledge.

(3) Boghossian (1989, pp. 17 ff.; cf. Goldman 1993, p. 25) has argued that privileged self-knowledge must involve '*cognitive achievement*' in the sense of requiring observation or the performance of some inference based on some observation. Self-knowledge is a cognitive achievement because it is subject to direction, is fallible, and incomplete. According to Boghossian, the inclusion theory of privileged self-knowledge cannot account for any of these characteristics and therefore is cognitively insubstantial; it is sham knowledge. The reason it cannot account for these characteristics is that it relies on reliabilism according to which the justifi-

[9] Cf. Cohen (1986), Dretske (1970), Lewis (1979). Reliabilism and contextualism are compatible positions.

cation condition for knowledge is replaced by some non-normative notion of evidential support. Boghossian compares Burge's analysis of judgments like "I am thinking that water is wet" to the analysis of logical truths such as "I am here now". Whenever one thinks "I am here now" one thinks a true thought of which one knows that it is true. The judgment is justified not because one possesses special knowledge about oneself, one's location, and time but simply in virtue of the meanings of the indexical elements involved. Mastering the words 'I', 'here', and 'now' is sufficient for knowing that "I am here now" is true. In fact, Burge asserts that *whenever* one thinks about one's occurrent thoughts one is epistemically justified in making a judgment of the kind "I am thinking that P". It is not necessary to ground self-reflexive thoughts on anything else such as other beliefs or observations. Burge writes:

> 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. [...] Justification lies not in the having of supplemental background knowledge, but in the character and function of the self-evaluating judgments (1988, p. 660).

It is the non-normative notion of evidential support Boghossian quarrels with. A reliabilist notion of self-knowledge, he argues is "based on nothing – at any rate, on nothing empirical" (1989, p. 17). For self-knowledge to be genuine knowledge, it must be justified by observation or inference based on some observation.

This is the most basic of the three incompatibilist objections, since it presents a head-on attack on the epistemology underlying the inclusion theory of self-knowledge. One way to fend off Boghossian's charge would be to come up with a general criticism of the internalist justification condition for knowledge. However, I don't see this strategy getting us very far since a *general* dispute between internalism and reliabilism is unlikely to provide a knockdown argument for the reliabilist treatment of *self*-knowledge.[10] Fortunately, there is another way to demonstrate that the reliabi-

[10] It is worth noting that arguably the most powerful internalist objection to reliabilism may apply to knowledge of the external world but doesn't apply to introspective knowledge. The argument says that, according to the reliabilist criteria, a person may be highly irrational and irresponsible in accepting a belief, when judged in the light of his own subjective conception of the situation, and may still turn out to be epistemologically justified and to possess knowledge because the belief is reliably formed. To BonJour (1985, ch. 3) and others this seems highly counterintuitive: If one has good reason to think that one's

list construal of privileged self-knowledge is cognitively substantial. I will neutralize Boghossian's objection by showing that, despite its commitment to reliabilism, the inclusion theory of self-knowledge can account for the three qualities he claims are distinctive of genuine self-knowledge, namely *directability*, *fallibility*, and *incompleteness*.

Directability: One of the essential properties of self-knowledge, so Boghossian says, is that it is subject to direction: "how much you know about your thoughts should [...] depend on how much *attention* you are paying to them" (1989, p. 19). The inclusion theory, he claims, is incapable of accounting for this feature for, on this picture, "you do not know your thoughts on the basis of evidence". This is false. Nothing prevents the inclusion theory from acknowledging that one can attend to and investigate one's mental states with varying degree of care. Frequently one has to take a second 'look' to find out, say, that the pain one is feeling is of the throbbing rather than the stinging or grinding kind. The inclusion theory can account for this fact in the following way: By attending to one's thought that P one can become a reliable indicator not only of P (pain) but of specific forms of P (throbbing pain). What the inclusion theory denies, however, is that by paying closer attention to one's thought that P one's knowledge of this thought somehow becomes more reliable. Like any other kind of factual knowledge, privileged self-knowledge is an absolute concept. Either I know that I am thinking P or

---

belief isn't reliable, then, even if it is reliable, it doesn't qualify as knowledge. Now this argument doesn't undermine reliabilism regarding privileged self-knowledge. The crucial question is: What could count as reasons for thinking that one's introspective beliefs are unreliable? There are two obvious candidates for such evidence – behavioral evidence and findings of a brain-scanning device. Behavioral evidence doesn't do the job since it is of empirical origin. For the above anti-reliabilist argument to be applicable to privileged self-knowledge, one's reason for thinking that one's second-order belief is unreliable would to be based on non-empirical origin. For the above anti-reliabilist argument to be applicable to privileged self-knowledge, one's reason for thinking that one's second-order belief is unreliable would have to be based on non-empirical evidence. The second candidate, a brain-scanning device, fails to complete the anti-reliabilist argument since the postulation of such a device doesn't hold up to scrutiny (cf. Bernecker 1996a, pp. 127–34; Shoemaker 1994, pp. 249–90). Hence, there is no way to make sense of the idea that one has reason to doubt the reliability of one's introspective beliefs about one's occurrent thoughts. Moreover, given the coherence theory of justification (and other inferential conceptions of self-knowledge), it is not only unreasonable but even impossible to have reason to question the reliability of all of one's self-referential beliefs. On this view, an introspective belief is justified by reference to other introspective beliefs. General doubts concerning the justificatory status of one's introspective beliefs would lead to a vicious regress (cf. Boghossian 1989, pp. 8–11; BonJour 1985, p. 51).

I don't. There is nothing like knowing it better. From an epistemic point of view, my belief that I am in some kind of pain or other is just as good as my belief that I am suffering from intense throbbing pain. The second kind of belief might of course prove to be more useful as it allows a physician to develop a more precise diagnosis of my ailment. But this is beside the point, since whether or not a belief is useful isn't relevant for it qualifying as knowledge.

Fallibility and *Incompleteness*: "The most important consideration", according to Boghossian, "against an insubstantial construal of self-knowledge" is that it cannot account for the fallibility and incompleteness of self-knowledge: "mental [...] events may occur of which one remains ignorant; and [...] even when one becomes aware of an event's existence, one may yet misconstrue its character, believing it to have a property it does not in fact possess" (1989, p. 19). Boghossian doesn't explain why fallibility and incompleteness are supposed to be necessary conditions of knowledge. Presumably he holds that the very concept of knowledge implies the possibility of mistake, and that when there is no possibility of getting things wrong, all talk of knowledge is out of place. Now, I don't see any reason why knowledge cannot be complete and unfailing. But quite apart from this, Boghossian errs in maintaining that the inclusion theory implies that privileged self-knowledge is infallible and complete. This error goes back to Burge himself who holds that the inclusion theory renders self-knowledge "self-verifying (hence infallible)" (1988, p. 658n). Given the inclusion theory, it is true that one cannot be wrong about the fact that one is thinking of P (e.g., water) rather than Q (e.g., twater). However, the inclusion theory is consistent with our ignorance and fallibility regarding other aspects of our mental condition. This needs further explanation.

In (1996b) I have argued that though the inclusion theory explains knowing it is P I am thinking about, it doesn't explain how I can have privileged self-knowledge that the state I occupy is a state of believing rather than, say, a state of doubting, or expecting etc. In other words, the inclusion theory explains privileged access to one's *contents* but not to one's *attitudes*. The reason is that a state's attitude isn't part of its content and it is only the content that is automatically included in the that-clause of a self-referential cognition. Since self-knowledge consists in the identification of the attitude as well as the content, the inclusion theory doesn't provide a complete account of privileged self-knowledge. Moreover given externalism, self-knowledge of the attitudinal component is vulnerable to

brute error and doesn't have the same kind of privilege as self-knowledge of content. The suggestion is that because of external determination of attitude concepts such as 'to believe', 'to doubt', and 'to expect', empirical investigation might be needed to know the mode in which one's thought content is realized. The inclusion theory therefore cannot be extended to provide a solution to privileged self-knowledge of attitudinal components, and that there is no indication that there is some other externalist account to be had of this kind of knowledge. Thus, there is at least one respect in which the inclusion theory can allow for privileged self-knowledge to be fallible and incomplete.

Now, Boghossian could turn around and claim that the point about attitude-identification, rather than lending support to the inclusion theory of self-knowledge by rendering it cognitively substantial, works to its disadvantage. The fact that the inclusion theory cannot explain privileged access to one's attitudes doesn't point to a limitation of our capacity to know our own minds, instead it reveals a limitation on the part of the inclusion theory. The notion of self-knowledge underlying the inclusion theory is sham knowledge because it cannot account for privileged access to one's attitudes. Without being able to respond to this challenge in detail, I want to suggest that it is only through the study of externalism that a reasonable notion of self-knowledge emerges. Privileged access to the attitudinal components of one's thoughts is one of the many Cartesian superstitions that the inclusion theory forces us to abandon. But even if it should turn out that attitude-identification is an essential aspect of first-person authority, the particular argument for incompatibilism presented in the beginning of the paper is flawed: Given both reliabilism and the relevant alternative account of knowledge, twinning and switching scenarios pose no problem for the doctrine of privileged self-knowledge of thought content. Moreover, nothing speaks against using reliabilism and the relevant alternative account to analyze privileged self-knowledge. We can conclude that externalism and privileged self-knowledge are consistent.[11]

[11] The preparation of this chapter was in part supported by a fellowship from the Stanford Humanities Center. For valuable comments on earlier drafts of this paper, I am grateful to Tony Brueckner, Fred Dretske, Frank Hofmann, Peter Ludlow, Carlos Moya, John Perry and Ken Taylor.

REFERENCES

Bernecker, S. 1996a. Davidson on First-Person Authority and Externalism. *Inquiry* 39: 121–39.

———. 1996b. Externalism and the Attitudinal Component of Self-Knowledge. *Nous* 30: 262–75.

Boghossian, P. 1989. Content and Self-Knowledge. *Philosophical Topics* 17: 5–26. Chapter 6 in this volume.

———. 1992: Externalism and Inference. *Philosophical Issues* 2: 11–28.

———. 1994. The Transparency of Mental Content. *Philosophical Perspectives* 8: 33–50.

BonJour, L. 1985. *The Structure of Empirical Knowledge.* Cambridge, MA: Harvard University Press.

———. 1991. Is Thought a Symbolic Process? *Synthese* 89: 331–52.

Brueckner, A. 1990. Scepticism about Knowledge of Content. *Mind* 99: 447–51.

———. 1992. Semantic Answers to Skepticism. *Pacific Philosophical Quarterly* 73: 200–19.

———. 1994. Knowledge of Content and Knowledge of the World. *Philosophical Review* 103: 107–37.

———. 1997. Externalism and Memory. *Pacific Philosophical Quarterly* 78: 1–12. Chapter 19 in this volume.

Burge, T. 1988. Individualism and Self-Knowledge. *Journal of Philosophy* 85: 649–63. Chapter 4 in this volume.

———. 1996. Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society* 117: 91–116. Chapter 15 in this volume.

———. 1998. Memory and Self-Knowledge. Chapter 21 in this volume.

Cohen, S. 1986. Knowledge and Context. *Journal of Philosophy* 83: 574–83.

Dretske, F. 1970. Epistemic Operators. *Journal of Philosophy* 67: 1007–23.

———. 1981. The Pragmatic Dimension of Knowledge. *Philosophical Studies* 40: 363–78.

Goldman, A. 1976. Discrimination and Perceptual Knowledge. *Journal of Philosophy* 73: 771–91.

———. 1993. The Psychology of Folk Psychology. *Behavioral and Brain Sciences* 16: 15–28.

Hofmann, F. Mscr.: Externalism and Memory. University of Tübingen, Germany.

Hooker, M. 1973. In Defense of the Principle for Deducibility of Justification. *Philosophical Studies* 24: 402–6.

Lewis, D. 1979. Scorekeeping in a Language Game. *Journal of Philosophical Logic* 8: 513–43.

Ludlow, P. 1995a. Externalism, Self-Knowledge, and the Prevalence of Slow Switching. *Analysis* 55: 45–9. Chapter 12 in this volume.

———. 1995b. Social Externalism, Self-Knowledge, and Memory. *Analysis* 55: 157–9. Chapter 17 in this volume.

———. 1996. Social Externalism and Memory: A Problem?. *Acta Analytica* 14: 69–76. Chapter 18 in this volume.

McGinn, C. 1989. *Mental Content.* Oxford: Basil Blackwell.

Owens, J. 1992. Psychological Supervenience. *Synthese* 90: 89–117.

Shoemaker, S. 1994. Self-Knowledge and 'Inner Sense'. *Philosophy and Phenomenological Studies* 54: 249–314.

Warfield, T. 1992. Privileged Self-Knowledge and Externalism are Compatible. *Analysis* 52: 232–7. Chapter 11 in this volume.