PHILOSOPHICAL TOPICS
VOLUME XVII, NO.1, SPRING 1989

Content and Self-Knowledge

Paul A. Boghossian
University of Michigan

INTRODUCTION

1. This paper argues that, given a certain apparently inevitable thesis about content, we could not know our own minds. The thesis is that the content of a thought is determined by its relational properties.

The problem can be stated roughly, but intuitively, like this. We sometimes know our thoughts directly, without the benefit of inference from other beliefs. (Indeed, given a plausible internalism about justification, this claim is not merely true but necessary.) This implies that we know our thoughts either on the basis of some form of inner observation, or on the basis of nothing. But there is a difficulty either way. On the one hand, given that the content properties of thoughts are individuated in terms of their relational properties, we could not know what we think merely by looking inwards. What we would need to see, if we are to know by mere looking, is not there to be seen. And, on the other, there appear to be serious objections to the suggestion that we may know our thoughts on the basis of nothing.[1]

The paper proceeds as follows. Part I explains why we could not know our thoughts on the basis of reasoning or inference. Part II explains why we could not know them on the basis of looking. And Part III explains why we could not know them on the basis of nothing.

I consider the skeptical claim about self-knowledge to have the status of a paradox: apparently acceptable premises lead to an unacceptable con-

clusion. For I do not seriously envisage that we do not know our own minds. Our capacity for self-knowledge is not an optional component of our ordinary self-conception, a thesis we may be able to discard while preserving all that really matters. It is a fundamental part of that conception, presupposed by some of the very concepts that constitute it (consider intentional action). So long as we are not able to see our way clear to abandoning that conception—and I am assuming that we have not yet been shown how to do so—there can be no question of accepting the skeptical claim.[2]

The point of advancing it, then, is not to promote skepticism but understanding: I hope that by getting clear on the conditions under which self-knowledge is not possible, we shall better understand the conditions under which it is. I have to confess, however, that at the present time I am unable to see what those conditions might be.

A couple of preliminary remarks before we proceed. First, I propose to be reasonably serious in the use of the term 'knowledge': by 'self-knowledge' I shall mean not just a *true* belief about one's own thoughts, but a *justified* one. (I do not, however, propose to be so serious as to worry about the complexities induced by Gettier-style counterexamples.) Second, I want to keep the discussion as free as possible of problematic auxiliary assumptions about the nature of thought. In particular, I do not want to assume a 'language of thought' model of thinking. I hope one of these days to write a paper entitled "The Language of Thought Hypothesis in the Philosophy of Mind." It would argue that, contrary to what many people seem to believe, a language of thought model has profound and unexpected implications for the way we think about most mental phenomena. Issues about self-knowledge, in particular, are transformed by its assumption. The reason should be evident: a language of thought model implies that there are *type-type* correlations between certain purely formal and intrinsic properties of thoughts and their semantic properties. This is a heady assumption that stands to profoundly affect the account we are able to give of our capacity to know the semantic properties of thoughts. Too heady, I think, to be assumed uncritically and, hence, too heady for the purposes of this paper.

I

THE CHARACTER OF SELF-KNOWLEDGE

INFERENCE AND SELF-KNOWLEDGE 2. Many extravagant claims have been made about our capacity to know our own minds. Descartes, who was responsible for the worst excesses, taught many subsequent generations of philosophers that self-knowledge was both infallible and exhaustive. In contrast with our knowledge of other people's minds, Descartes held, our

access to our own contemporaneous mental states and events could issue neither in false belief nor in ignorance.

These famous Cartesian claims are not, of course, wholly without substance; for a certain restricted class of mental events—namely, sensations—they may even be true. For it does seem constitutive of, say, an occurrence of pain, that it register with us precisely as an occurrence of pain. And so, it seems not conceivable, in respect of facts about pain, that we should be either ignorant of their existence or mistaken about their character, just as the Cartesian doctrine requires.

But the corresponding theses about contentful or representational states carry little contemporary conviction. That we harbor a multitude of thoughts of whose existence we are unaware is a presupposition not only of Freudian theory, but of much of present-day cognitive science. And phenomena that are intelligible only if infallibility is false—self-deception, for instance—seem pervasive.

3. A Cartesian account, then, of the distinction between first-person and third-person knowledge of mind must be rejected. But we should be wary, in correcting for Cartesian excess, of recoiling too far in the opposite direction. For there remains, even after we have discarded the problematic Cartesian claims, a profound asymmetry between the way in which I know my own thoughts and the way in which I may know the thoughts of others. The difference turns not on the epistemic status of the respective beliefs, but on the manner in which they are arrived at, or justified. In the case of others, I have no choice but to *infer* what they think from observations about what they do or say. In my own case, by contrast, inference is neither required nor relevant. Normally, I know what I think—what I believe, desire, hope or expect—without appeal to supplementary evidence. Even where such evidence is available, I do not consult it. I know what I think directly. I do not defend my self-attributions; nor does it normally make sense to ask me to do so.[3]

Ryle attempted to deny all this.[4] He tried to defend the view that there is no asymmetry between first-person and third-person access to mental states. In both cases, he maintained, the process is essentially the same: ordinary inspection of ordinary behavior gives rise to the discovery of patterns in that behavior, which in turn leads to the imputation of the appropriate propositional attitudes.

The claim carries no conviction whatever. The trouble is not merely that it runs counter to all the relevant appearances, offering an implausible explanation for the knowledge we have of our thoughts. The trouble is that, for much of what we do know about our own thoughts, it can offer no explanation at all. Consider an act of entertaining a particular proposition. You think: Even lousy composers sometimes write great arias. And you know, immediately on thinking it, that that is what you thought. What explanation

can the Rylean offer of this? The difficulty is not merely that, contrary to appearance and the canons of epistemic practice, he has to construe the knowledge as inferential. The difficulty is that he has to construe it as involving inference from premises about behavior that you could not possibly possess. Your knowledge of that occurrent thought could not have been inferred from any premises about your behavior because that thought could not yet have come to have any traction on your behavior. So it's not merely that, on the Rylean view, you would have to know inferentially what you appear to know non-inferentially. It's that you would not know at all what you seem to know unproblematically.

Any inferential conception is likely to succumb to this sort of objection. Since the epistemic norms governing ascriptions of self-knowledge do not require possession of supplementary evidence, for any item of evidence insisted on by an inference-based account—whether it involve behavior or the environment or even the causal properties of thoughts—it should be possible to describe a situation in which you know your thoughts but you do not know the item in question.

INTERNALISM AND INFERENTIAL SELF-KNOWLEDGE 4. It is, actually, surprisingly little noticed that on an *internalist* conception of justification—a conception to which many philosophers remain profoundly sympathetic—knowledge of one's own mental states *has* to be non-inferential. On this view, the alternative is not merely implausible; it is incoherent. I shall explain.

The intuition that fuels internalism in the theory of justification is the thought that someone cannot count as justified in holding a certain belief if, judged from the standpoint of his own subjective conception of the situation, he may appear epistemologically irresponsible or irrational in accepting that belief. The intuition is effectively triggered by various examples in which, although a person's belief satisfies basic externalist demands—the belief is formed by a reliable belief-forming mechanism and so on—the person does not count as epistemically justified because, as far as he is concerned, he has no reason for accepting the belief and may, indeed, have reasons for rejecting it.

Consider Sam.[5] Sam believes himself to have the power of clairvoyance, though he has no reason for the belief, and some evidence—in the form of apparently cogent scientific results—against it. One day he comes to believe, for no apparent reason, that the President is in New York City. He maintains this belief, appealing to his alleged clairvoyant power, even though he is at the same time aware of a massive amount of apparently cogent evidence, consisting of news reports, allegedly live television pictures, and so on, indicating that the President is at that time in Washington, D.C. Now the President is in fact in New York City, the evidence to the

contrary being part of an official hoax. Moreover, Sam does in fact have completely reliable clairvoyant power under the conditions then satisfied, and his belief about the President did result from the operation of that power.

Is Sam justified in his belief about the President? Basic reliabilist demands are met; but the intuition persists that the belief cannot be epistemically justified because, judged from the standpoint of Sam himself, it is epistemically thoroughly irrational.

Examples such as this are at the heart of internalist dissatisfaction with externalist conceptions of justification; they motivate the requirement that, if a belief depends upon evidence, "the knower [must] grasp the connection between the evidence and what it is evidence for," if his belief is to be justified.[6]

Suppose, then, that the proposition that p depends on the proposition that q. According to internalism, if I am to be justified in believing that p, I must believe that p as a result both of my recognition that I believe that q, and that a belief that q justifies a belief that p. Spelling this out in explicit detail, we have:

1. I believe that p.
2. I believe that q.
3. The proposition that q justifies the proposition that p.
4. I know that I believe that q.
5. I know that a belief that q justifies a belief that p.
6. I believe that p as a result of the knowledge expressed in 4 and 5.

5. Now, there is, of course, a *standard* problem in holding that all knowledge of empirical propositions is inferential, that all beliefs can be justified only by reference to other beliefs. This is the problem of the regress of justification: If the belief that p is to count as justified, then the belief that q on which its justification depends must itself be justified. But if *all* beliefs can be justified only by reference to other beliefs, then the belief that q must itself be justified by reference to other beliefs. And this threatens to lapse into a vicious regress.

Any theory of justification must confront this problem. The available non-skeptical options—Foundationalism and Coherentism—are well known and need not be rehearsed here.

The point I wish to make, however, is that there is a *special* problem sustaining a thoroughly inferential conception of *self-knowledge,* one that is independent of the *standard* problem of the regress of justification.

6. In order to bring it out, waive the standard problem: let us not require that if the non-intrinsically credible proposition that p is to be justifiably believed, then it has to rest on a belief that q that is itself justified; let us simply require that the belief that p be justified *relative* to the belief that q, in accordance with standard internalist requirements.

Where the subject matter of concern is knowledge of one's own beliefs, the belief that p will be a belief to the effect that I have a certain belief, say, that I believe that r. Since we are supposing that all self-knowledge is inferential, there must be a belief on which this belief rests. Let that be the belief that s. Now, what would have to be true if I am to be justified in believing that I believe that r?

Taking into account the fact that the belief in question is a belief concerning my own beliefs, the conditions that would have to be satisfied, if I am to be justified in believing that I believe that r, are these:

1′. I believe that I believe that r.

2′. I believe that s.

3′. The proposition that s justifies the proposition that I believe that r.

4′. I know that I believe that s.

5′. I know that a belief that s justifies the belief that I believe that r.

6′. I believe that I believe that r as a result of the knowledge expressed in 4′ and 5′.

The problem is transparent. In order to be justified in believing that I have a certain belief, I must already know that I have some other belief (4′): In order to know that I believe that r, I must antecedently know that I believe that s. But how was knowledge of *this* belief acquired? On the assumption that all self-knowledge is inferential, it could have been acquired only by inference from yet other known beliefs. And now we are off on a vicious regress.[7]

The problem with sustaining a thoroughly inferential conception of *self-knowledge* should have been evident from the start. For the ordinary notion of being justified in believing a non-intrinsically credible empirical proposition *presupposes* self-knowledge. For it presupposes that one has grasped the fact that one's belief in that proposition bears some appropriate epistemic relation to one's other beliefs. In ordinary epistemological discussions this does not emerge as a problem because those discussions tend to focus exclusively on the justification of belief concerning the external world; they tend, understandably, to take knowledge of the beliefs themselves for granted. When such knowledge is not taken for granted, however, it emerges very clearly that not *all* knowledge of one's beliefs can be inferential. On pain of a vicious regress, it must be possible to know the content of some mental states non-inferentially.[8]

7. The intuitive epistemic facts indicate that knowledge of one's mental states is direct. And an internalist conception of justification implies that it has to be. There are two ways to accommodate this claim.

We may conclude, on the one hand, that self-knowledge is not in-

ferential because it is based on some form of inner observation; or, on the other, that it is not inferential because it is based on nothing—at any rate, on nothing empirical.

How might a contingent fact be known on the basis of nothing empirical? We shall consider that question in part III. Before that, however, I want to turn to asking whether we could know our thoughts on the basis of inner observation.

II

CONTENT AND KNOWLEDGE OF CONTENT

8. The suggestion that I know about my thoughts by being introspectively aware of them seems, from a phenomenological standpoint anyway, overwhelmingly plausible. It is not simply that I have reliable beliefs about my thoughts. I catch some of my thoughts in the act of being thought. I think: If she says that one more time, I'm leaving. And I am aware, immediately on thinking it, that that is what I thought. Can 'inner awareness' provide the right explanation for how I know my thoughts?

There are many aspects to this question. An exhaustive treatment would distinguish carefully between occurrent events—fleeting thoughts, sudden fancies—and standing states—fixed beliefs, stable desires—and would worry about the epistemological ramifications of that distinction. It would also distinguish between the distinct attitudes that one may sustain toward a given content—judging, believing, desiring, entertaining—and explore any corresponding epistemic differences. Here, however, I shall not be concerned with these important nuances. For my worry is that, given certain currently prevailing orthodoxies about content, it is impossible to see how *any* contentful state could be known on the basis of inner observation.

The difficulty stems from the contemporary commitment to a relationist conception of content: the view that the content properties of mental states and events are determined by, or supervenient upon, their *relational* properties. Intuitively, the difficulty 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? Articulating the intuitive problem in explicit detail is the task of the present part.

ANTI-INDIVIDUALISM AND SELF-KNOWLEDGE 9. The commitment to relationism is evident, of course, in *wide* or *anti-individualistic* conceptions of thought content. According to such views, many of a person's thought contents are necessarily dependent on relations that that person bears to the physical or, in some cases, social environment. The view is

supported by a series of now-famous thought experiments. Their strategy is to show that two individuals who are molecule-for-molecule duplicates of each other, may nevertheless think different thoughts if their environments differ from each other in certain specified ways. Thus, Putnam has argued that *part* of what makes it true that some of my thoughts involve the concept *water,* is that it is typically *in re* H_20 that I token those thoughts; a duplicate of mine, who grew up in an indistinguishably similar environment except that in it the liquid that filled the lakes and swimming pools consisted of XYZ and not H_2O, would not have the concept *water* but some other concept, *twater.* Similarly, Tyler Burge has argued that part of what makes it true that some of my thoughts involve the concept *arthritis* is that I live in, and defer to, a community in which the concept of arthritis is used in a certain way; a duplicate of mine, who grew up in an indistinguishably similar community except that in it the use of the concept was extended so as to cover all rheumatoid ailments, would not have the concept *arthritis* but some other concept, *tharthritis.*

10. Now, doesn't it follow from such anti-individualistic views that we cannot know our thoughts in a direct, purely observational manner? The following line of reasoning might seem to lead rather swiftly to that conclusion. To know my water thoughts, I would have to know that they involve the concept *water* and not the concept *twater.* But I could not know whether my thought involves the concept *water* or the concept *twater* without investigating my environment. For what I would need to know is whether it was typically *in re* H_2O or typically *in re* XYZ that I token my thoughts; and I certainly would have to investigate my environment in order to know that. I could hardly know such facts by mere introspection. It would seem to follow, therefore, that I could not know the contents of my thought purely observationally: I would have to *infer* what I think from facts about my environment.

This line of reasoning is no doubt too swift. As it stands it appears to be making problematic assumptions about the conditions required for knowledge.[9] Consider perceptual knowledge. Someone may know, by looking, that he has a dime in his hand. But it is controversial, to put it mildly, whether he needs to know all the conditions that make such knowledge possible. He need not have checked, for example, that there is no counterfeit money in the vicinity, nor does he need to be able to tell the difference between a genuine dime and every imaginable counterfeit that could have been substituted. The ordinary concept of knowledge appears to call for no more than the exclusion of "relevant" alternative hypotheses (however exactly that is to be understood); and mere logical possibility does not confer such relevance.

Similar remarks apply to the case of self-knowledge. And so, since under normal circumstances the *twater* hypothesis is not a relevant alter-

native, we ought not to assume, as the swift argument evidently does, that we could not know our actual thought contents unless we are able to discriminate between them and their various twin counterparts.

II. The swift argument, however, suggests a slower and more convincing argument for the same conclusion. For it seems fairly easy to describe scenarios in which the twin hypotheses *are* relevant alternatives, but in which they are, nevertheless, not discriminable non-inferentially from their actual counterparts.

Imagine that twin-earth actually exists and that, without being aware of it, S undergoes a series of switches between earth and twin-earth. Most anti-individualists agree that, if a person were to remain in each situation long enough, that person would eventually acquire the concepts appropriate to that situation.[10] There are two ways to imagine the final outcome. On the one hand, we may imagine that after a series of such switches, S ends up with *both* earthian and twin-earthian concepts: thoughts involving both *arthritis* and *tharthritis* are available to him. Or, alternatively, we may imagine that with every such slow switch a wholesale displacement of S's resident concepts takes place, so that at any given time either the earthian or the twin-earthian concepts are available to him, but not both.

The story is usually told, I believe, in the second of these two ways; though so far as I can tell, it is perfectly coherent—and a lot more interesting—to tell it the other way. Still, in the interests of keeping matters as simple as possible, I shall follow tradition and imagine only the second version.[11] I invite you to consider, then, a thinker S, who, quite unawares, has been shuttled back and forth between earth and twin-earth, each time staying long enough to acquire the concepts appropriate to his current situation, and at the expense of the concepts appropriate to his previous situation.

What does S know? By assumption, he is not aware that the switches have taken place and nothing about his qualitative mental life or his perceived environment tips him off. Indeed, S may not even be aware of the existence of twin-earth or of the dependence of content on environment. As far as S is concerned, he has always lived on earth. If someone were to ask him, just after one set of twin-earthian concepts has been displaced by a set of earthian ones, whether he has recently thought thoughts involving an arthritis-like concept distinct from *arthritis,* S would presumably say "no." And yet, of course, according to the anti-individualist story, he has. His knowledge of his own past thoughts seems very poor, but not presumably because he simply can't *remember* them. Could it be because he never knew them?

Let us in fact confront that question directly. Does S know what he is thinking while he is thinking it? Suppose he is on twin-earth and thinks a thought that he would express with the words "I have arthritis." Could he

know what he thought? The point to bear in mind is that the hypothesis that he thought *I have arthritis* is now a relevant alternative. He, of course, is not aware of that, but that doesn't change matters. Epistemic relevance is not a subjective concept. 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. Similarly, 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?[12]

INDIVIDUALIST CONTENT AND SELF-KNOWLEDGE

12. Ever since Putnam first invented twin-earth, philosophers have expressed concern about the compatibility of wide individuation with the direct character of self-knowledge.[13] In the previous section I have tried to show that these concerns are in order, that there is indeed a problem reconciling the thesis with the intuitive facts. As I shall try now to explain, however, the problem about self-knowledge was there all along and the recent emphasis on *widely* individuated content betrays a misunderstanding: even if no external factors were involved in fixing mental content, on any currently acceptable account of the internal determinants, the difficulty about self-knowledge would still remain.

The point is that according to currently prevailing orthodoxy, even the internal (or narrow) determinants of a mental event's content are relational properties of that *event* (although they are, of course, intrinsic properties of the *thinker* in whom the events occur.)

An example of a properly non-relationist conception of content is provided by the imagistic theory of the British Empiricists. According to this theory, thinking the thought that p involves entertaining an image that represents that p. And the facts in virtue of which an image represents a particular state of affairs are said to depend exclusively on the intrinsic properties of the image. Neither tenet is considered plausible today. Thinkings are not imagings; and, in any case, the representational properties of images are not determined by their intrinsic properties.

Indeed, according to contemporary conviction, there is *no* property intrinsic to a mental event—certainly no *naturalistic* intrinsic property—that could serve as the complete determinant of that event's representational content. In effect, the only idea around about what narrow properties of an event might fix its content is the suggestion that it is some subset of

the event's *causal* properties. The central functionalist idea here is that the content of a mental event is determined by that event's causal role in reasoning and deliberation and, in general, in the way the event interacts with other events so as to mediate between sensory inputs and behavioral outputs. On the assumption, then, that no external factors are involved in content individuation, the facts in virtue of which a thought is a thought about *water,* as opposed to a thought about *gin,* have to do with the thought's causal properties: thoughts with causal role R are thoughts about water, whereas thoughts with causal role R′ are thoughts about gin.

Consider now a particular episode of thinking *water is wet.* How, on the dominant functionalist picture, might I know that that is what I thought? To know that I just had a *water* thought, as opposed to a *gin* thought (which, unlike a *twater* thought counts as a relevant alternative even in the absence of special circumstances) I would have to know, it seems, that my thought has the causal role constitutive of a *water* thought, as opposed to one constitutive of a *gin* thought. But it doesn't seem possible to know a thought's causal role directly. The point derives from Hume's observation that it is not possible to ascertain an item's causal properties noninferentially, by mere inspection of its intrinsic properties; discovering them requires observation of the item's behavior over time.

But, again, this would appear to imply that I would have to *reason* my way to a thought's content; and reason to it, moreover, from facts about its causal role that I do not necessarily possess. How, then, could I know my thoughts at all?—much less know them directly?

KNOWLEDGE OF RELATIONS 13. It might be suggested that the appearance of a difficulty here is being generated by appeal to a false principle: namely, that in order to know a mental event one must know how things stand with respect to the conditions that individuate that event.[14]

The cogency of the argument would certainly be at risk if such a principle were being assumed. For it is clearly not in general true that to know whether an object x has a property P one has to know how things stand with respect to the facts on which P supervenes. For example, the roundness of this coin in my hand supervenes on a mass of facts concerning the arrangement of molecules at its boundary; but I do not need to know those facts in order to know that the coin is round.

It is fortunate for my argument, therefore, that it assumes no such principle. What it does assume is different and considerably more plausible. Namely this: That you cannot tell by mere inspection of an object that it has a given *relational* or *extrinsic* property. This principle is backed up by appeal to the following two claims, both of which strike me as uncontestable. That you cannot know that an object has a given rela-

tional property merely by knowing about its *intrinsic* properties. And that mere inspection of an object gives you at most knowledge of its intrinsic properties.

Uncontestable or not, it may yet seem that there are exceptions to the principle that an extrinsic property can never be detected by mere inspection.

Consider monetary value. Being a dime is not an intrinsic property of an object: for something to be a dime it must bear a number of complicated relations to its economic and social environment. And yet, we seem often able to tell that something is a dime purely observationally, by mere inspection of its intrinsic properties. Counterexample.

Not quite. The reason an extrinsic property seems, in this case, ascertainable by mere inspection, is due to the fact that possession of that property is correlated with possession of an intrinsic property that is ascertainable by mere inspection. The reason that the coin's dimehood seems detectable by mere inspection derives from the fact that its having the value in question is neatly encoded in several of its purely intrinsic properties: in the phrase "ten cents" that is inscribed on it, and in several other of its size, shape, and design characteristics.

To see clearly that it is only because of this feature that we are able to "inspect" the coin's value properties, consider a monetary system in which *all* coins, regardless of value, share their intrinsic properties: they are all minted of precisely the same metal, are all precisely of the same shape, size, and design. As far as their intrinsic properties are concerned, nothing serves to distinguish between coins of different value. Nevertheless, the coins are not all of equal value; and, let us suppose, what determines a coin's value is the mint it was minted at: coins minted at "five cent" mints are worth five cents, those at "ten cent" mints, ten cents, and so on. It should be obvious that the value of *these* coins is not ascertainable by mere inspection; one would have to know something about their historical properties.

If this is right, it shows that our normal ability to "inspect" monetary value cannot help explain our ability to know our thought contents directly. First, because the feature that helps explain our knowledge in the former case—the correlation between the coin's monetary value and possession of certain intrinsic properties—does not obtain in the latter: facts about a thought token's content are not correlated with any of that token's purely intrinsic properties.[15] And second, because even if this were not true, that would still not explain how we might know our thoughts directly. For the process by which we know the coin's value is not really inspection, it's inference: you have to deduce that the coin is worth ten cents from your knowledge of its intrinsic properties plus your knowledge of how those in-

trinsic properties are correlated with possession of monetary value. And our knowledge of thought is not like that.

III
IS SELF-KNOWLEDGE A COGNITIVE ACHIEVEMENT?

14. Many philosophers would agree, I think, with the conclusion of the previous part: that if we had to know our thoughts on the basis of inner observation, then we couldn't know our thoughts. It has certainly become very popular to claim that an observational model of self-knowledge is mistaken. Thus Burge:

> If one thinks of one's relation to the subject matter of basic self-knowledge on an analogy to one's relation to objects of empirical investigation, then the view that one's thoughts (the subject matter) are dependent for their natures on relations to the environment will make it appear that one's knowledge of one's thoughts cannot be any more direct or certain than one's knowledge of the environment. . . . This line of reasoning is deeply misconceived.[16]

Donald Davidson has sounded a similar theme:

> I can tell by examining my skin what my private or 'narrow' condition is, but nothing I can learn in this restricted realm will tell me that I am sunburned. The difference between referring to and thinking of water and referring to and thinking of twater is like the difference between being sunburned and one's skin being in exactly the same condition through another cause. The semantic difference lies in the outside world, beyond the reach of subjective or sublunar knowledge. So the argument might run.
>
> This analogy, between the limited view of the skin doctor and the tunnel vision of the mind's eye, is fundamentally flawed.[17]

But it is not as if, in opposing an observational model of self-knowledge, these philosophers are suggesting that knowledge of thought is inferential. The claim is, rather, that the correct way to explain the direct and authoritative character of self-knowledge is to think of it as based on nothing—at any rate, on nothing empirical.

COGNITIVELY INSUBSTANTIAL JUDGMENTS Ordinarily, to know some contingent proposition you need either to make some observation, or to perform some inference based on some observation. In this sense, we may say that ordinary empirical knowledge is always a *cognitive achievement* and its epistemology always *substantial*. How could a judgment about a contingent matter of fact count as knowledge and yet not be a cognitive

achievement? Or, to put the question another way, how could a contingent proposition be known directly, and yet not through observation?

Consider the judgment *I am here now.* Any token of this contingent judgment would be true and justified. But, in contrast with ordinary empirical judgments, the thinker is not required to possess any evidence for his judgment; he needs only to think it. The judgment is true and justified as soon as thought. The thinker counts as knowing something thanks not to the possession of any empirical evidence on his part, but simply courtesy of the concepts involved.

Consider another example. Suppose that the Kantian thesis, that experience of the world as containing substances is a precondition for experiencing it at all, is correct. It would follow, on such a view, (and ignoring for present purposes the distinction between experiencing and knowing), that knowledge that the (experienced) world contains substances is knowledge that is cognitively insubstantial. To know the fact in question a thinker is not required to possess any particular item of empirical evidence; he needs merely to experience. The truth of, and warrant for, the belief are secured, not by evidence, but by the satisfaction of certain very general conditions on experience. The thinker counts as knowing something thanks not to the possession of any evidence on his part, but simply courtesy of those general facts.

A third example. According to some philosophers, certain self-regarding judgments are essentially self-verifying. Antecedent to the judgment that I am jealous, for example, there may be no fact of the matter about whether I am; but thinking it makes it so.[18] The judgment that I am jealous, when made, is, therefore, both true and justified. But, again, no evidence is required for the judgment. To know the fact in question, I am not required to possess any particular item of empirical evidence; I need merely to make the judgment. I count as knowing something thanks not to the possession of any evidence on my part, but simply courtesy of the self-verifying nature of the judgment involved.

These examples illustrate three different kinds of contingent judgment which one may be justified in making even in the absence of any empirical evidence. The warrant for such judgments derives from other sources: from the meanings of the concepts involved, or from the satisfaction of certain general conditions, or from the judgment-dependent character of the phenomena being judged. Whatever the source, no observation, or inference based on observational premises, is required or relevant. These judgments, when known, constitute knowledge that is based on nothing empirical. In my terms, they are not cognitive achievements and are subject, therefore, to an insubstantial epistemology.[19]

The relevance of such judgments ought to be clear. So long as knowledge of thought is construed as dependent on evidence, it seems impossible

to understand how we could know our thoughts. That is what the argument of the previous two parts amounts to. If, however, self-regarding judgments could be understood along cognitively insubstantial lines—as the sorts of judgment which, for one reason or another, might be known without empirical evidence—then we might be able to explain how we know our thoughts, consistent with the admission that we do not know them on the basis of observation, or of inference based on observation. Could self-knowledge be, in this way, cognitively insubstantial?

COGNITIVELY INSUBSTANTIAL SELF-KNOWLEDGE 15. It is hard to see how it could be. Knowledge that is not a cognitive achievement would be expected to exhibit certain characteristics—characteristics that are notably absent from self-knowledge. For instance, and unlike ordinary empirical knowledge, you would not expect cognitively insubstantial knowledge to be subject to direction: how much you know about your thoughts should not depend on how much *attention* you are paying to them, if you do not know your thoughts on the basis of evidence. And yet it does seem that, within bounds anyway, self-knowledge can be directed: one can decide how much attention to direct to one's thoughts or images, just as one can decide how much attention to pay to objects in one's visual field.[20]

Or consider the fact that some adults are better than others at reporting on their inner states; and that most adults are better than children. How is this to be explained if self-knowledge is not to be thought of as an information-sensitive capacity that may be subject to cultivation or neglect?

The most important consideration, however, against an insubstantial construal of self-knowledge derives not so much from these observations but from a claim they presuppose: namely, that self-knowledge is both fallible and incomplete. In both the domain of the mental and that of the physical, events may occur of which one remains ignorant; and, in both domains, 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. How is this to be explained? I know of no convincing alternative to the following style of explanation: the difference between getting it right and failing to do so (either through ignorance or through error) is the difference between being in an epistemically favorable position with respect to the subject matter in question—being in a position to garner the relevant evidence—and not. To put this point another way, it is only if we understand self-knowledge to be a cognitive achievement that we have any prospect of explaining its admitted shortcomings.

There is an irony in this, if it's true. Since Descartes, self-knowledge has been thought to present special philosophical problems precisely because it was held to be immune to cognitive deficit. The assumption was that we knew—or anyway had some idea—how to explain *imperfect* cog-

nitive mechanisms; what seemed to elude explanation was a cognitive faculty that never erred. This line of thought seems to me to be exactly backwards. If Descartes' hyperbolic claims were right—if self-knowledge really were immune to error and ignorance—the temptation to explain it in an epistemologically deflationary way would be overwhelming. As it is, however, the Cartesian claims are incorrect and the epistemology of self-knowledge, thereby, substantial.

16. Strange to discover, then, that deflationary accounts of self-knowledge appear to be gaining widespread acceptance.[21] I have already mentioned some general reasons for being suspicious of such accounts. In the remainder of this paper I propose to look at Burge's provocative proposal in detail, outlining the specific ways in which, as I see it, it fails as an account of self-knowledge.

BURGE: SELF-KNOWLEDGE AND SELF-VERIFICATION

17. According to Burge, it is a fundamental error to think that self-knowledge is a species of cognitive achievement. As he puts it, it is a mistake to think that, in order to know a thought, one must know a lot about it. Rather,

> [t]he source of our strong epistemic right, our justification, in our basic self-knowledge is not that we know a lot about each thought we have. . . . It is that we are in the position of thinking those thoughts in the second-order, self-verifying way.[22]

How is this to be understood?

Consider the following judgment about what I am thinking:

> I judge: I am thinking that writing requires concentration.

In such a judgment, Burge points out, the subject matter of the judgment is not merely contingently related to the thoughts one thinks about it. The judgment is self-referential and self-verifying. The second-order judgment to the effect that I'm thinking that writing requires concentration could not exist unless I were to think, through that very thought, that writing requires concentration. The thought I am making a judgment about is self-referentially fixed by the judgment itself; and the judgment is thereby self-verifying. At least in this sort of case, then, it appears that one need know nothing about a thought in order to know that one has thought it; one need only think the thought as part of a second-order thought that asserts its occurrence. Since such thoughts are, as Burge correctly points out, logically self-verifying, they are guaranteed to be true as soon as thought. Hence, they would appear to constitute authoritative and non-inferential knowledge of thought, the relational character of the properties that determine thought content notwithstanding.

Burge calls this sort of self-verifying, self-regarding judgment *basic self-knowledge*. Let us start with the following question: how much of di-

rect self-knowledge is basic self-knowledge? How well does Burge's paradigm explain the general phenomenon?

18. We may begin by noting that it does not at all explain our knowledge of our *standing* mental states. Judgments concerning such states, for example,

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. These self-regarding judgments do not conform to Burge's paradigm. This would appear to be a serious problem. After all, we do know about our beliefs and desires in a direct and authoritative manner, and Burge's proposal seems not to have the resources to explain how.

How does his proposal fare in connection with *occurrent* events? In this domain, too, its applicability seems rather limited. 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.

The best possible case for Burge's purposes will involve a self-regarding judgment about a mere thinking or entertaining of a proposition—a judgment of a the form

I judge: I think that writing requires concentration.

And even here, the judgment will only prove self-verifying if the time at which the judgment is made is *absolutely coincident* with the time at which the thought being judged about is thought. 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. It is only under this very special condition that the thinking of the proposition in question is presupposed by the very act of making a judgment about it; and, hence, only under this very special condition that the judgment is self-verifying.[23] If, for example, the judgment concerned an act of entertaining a proposition that preceded the act of making judgment by even the smallest interval of time, as in

I judge: I just now thought that writing requires concentration

then, since it need not be true that I had that thought *then* in order to make this judgment *now,* such a judgment would not be self-verifying and,

hence, would constitute a species of self-knowledge that is not subject to Burge's deflationary paradigm.

But is it not precisely knowledge of this form—knowledge of what one has thought immediately after one has thought it—that we think of as central to our capacity for self-knowledge? We are struck by our ability to know, non-inferentially and authoritatively, that a certain mental event has occurred, immediately on its having occurred. We think: Writing requires concentration. And then we know, directly and unproblematically, that that is what we thought. A first-order thought occurs. And we are then able, without the benefit of inference, to form a correct judgment about what thought that was. The second-order judgment in these central cases is not self-verifying. Such cases are not instances of "basic self-knowledge" in Burge's sense. How does his proposal help explain how they are possible? The fact that, *had* the thought been part of a second-order judgment, then that judgment would have been self-verifying, does not help explain how we are able to know what thought it was, given that it *wasn't* part of such a judgment. First-order thoughts that are not part of second-order thoughts are directly knowable. Arguably, acts of knowing such thoughts are paradigm cases of self-knowledge. And Burge's proposal seems incapable of explaining how they are possible.

19. Still, even if Burge's proposal does not explain the central cases, does it not supply us with at least *one* case in which a thought is known directly despite the relational nature of its individuation conditions? And isn't that enough to dislodge our intuition that relationism is irreconcilable with directness?

If Burge's self-verifying judgments were instances of genuine knowledge, then they would indeed dislodge the problematic intuition. But I am not convinced that they are.

Consider again the case of the person who undergoes a series of slow switches between earth and twin-earth. Burge observes:

> In the former situation, the person may think "I am thinking that water is a liquid." In the latter situation, the person may think "I am thinking that twater is a liquid." In both cases the person is right and as fully justified as ever. The fact that the person does not know that a switch has occurred is irrelevant to the truth and justified character of these judgments. Of course, the person may learn about the switches and ask "Was I thinking yesterday about water or twater?"—and yet not know the answer. Here knowing the answer may sometimes depend on knowing empirical background conditions. But such sophisticated questions about memory require a more complex story.[24]

These remarks strike me as puzzling. They amount to saying that, although S will not know tomorrow what he is thinking right now, he does know right now what he is thinking right now. For any given moment in the

present, say t1, S is in a position to think a self-verifying judgment about what he is thinking at t1. By Burge's criteria, therefore, he counts as having direct and authoritative knowledge at t1 of what he is thinking at that time. But it is quite clear that tomorrow he won't know what he thought at t1. No self-verifying judgment concerning his thought at t1 will be available to him then. Nor, it is perfectly clear, can he know by any other non-inferential means. To know what he thought at t1 he must discover what environment he was in at that time and how long he had been there. But there is a mystery here. For the following would appear to be a platitude about memory and knowledge: if S knows that p at t1, and if at (some later time) t2, S remembers everything S knew at t1, then S knows that p at t2. Now, let us ask: *why* does S not know today whether yesterday's thought was a *water* thought or a *twater* thought? The platitude insists that there are only two possible explanations: either S has forgotten or he *never* knew. But surely memory failure is not to the point. In discussing the epistemology of relationally individuated content, we ought to be able to exclude memory failure by stipulation. It is not as if thoughts with widely individuated contents might be easily known but difficult to remember. The only explanation, I venture to suggest, for why S will not know tomorrow what he is said to know today, is not that he has forgotten but that he never knew. Burge's self-verifying judgments do not constitute genuine knowledge. What other reason is there for why our slowly transported thinker will not know tomorrow what he is said to know directly and authoritatively today?[25]

In sum, Burge's self-verifying judgments seem to me neither to explain the central cases, nor to provide particularly compelling examples of special cases in which a relationally individuated thought is known non-inferentially.

CONCLUSION 20. In this paper, I have attempted to map out the available theoretical options concerning self-knowledge. And I have argued that none of the options work. It seems to me that we have a serious problem explaining our ability to know our thoughts, a problem that has perhaps not been sufficiently appreciated. As I said in the introduction, however, the point of the exercise is not to promote skepticism, but understanding. I am confident that one of the options will work; but I think we need to think a lot harder before we are in a position to say which one.[26]

NOTES

1. How a contingent proposition might be known on the basis of nothing will be explained in part III. A word also about 'inner observation': It makes no difference to the argument of this paper if you think of inner observation as amounting to traditional intro-

spection, or if you think of it as amounting to the operation of some Armstrong-style 'brain-scanner'. What *is* crucial to inner observation models of self-knowledge is the claim that beliefs about one's own thoughts are justified by the deliverances of some internal monitoring capacity, much like beliefs about the external environment are justified by the deliverances of an external monitoring capacity (perception).

For 'brain-scanners' see D. Armstrong, *A Materialist Theory of the Mind* (London: RKP, 1968). For a useful survey of various conceptions of introspection see William Lyons, *The Disappearance of Introspection* (Cambridge: MIT Press, 1986).

2. For arguments in support of the indispensability of the ordinary conception, see my "The Status of Content," forthcoming in *The Philosophical Review.*
3. Many philosophers have pointed this out. See, for example, D. Davidson, "Knowing One's Own Mind," *Proceedings of the American Philosophical Association* (1986): 441–442.
4. See Gilbert Ryle, *The Concept of Mind* (London: Hutchinson, 1949).
5. The example is adapted from L. Bonjour, *The Structure of Empirical Knowledge* (Cambridge: Harvard University Press, 1985), 38–40.
6. D. J. O'Connor and B. Carr, *Introduction to the Theory of Knowledge* (Minneapolis: University of Minnesota Press, 1982), 75. Of course, there are responses available to the externalist. I am not going to consider them here because I am not here trying to *argue* for internalism; I am just describing it. Again, for detailed discussion see Bonjour, *op. cit.*
7. The regress does not particularly depend on the fact that the relation between the beliefs consists in *inference*. There are possible coherentist views according to which mediated justification consists not in inference but in "membership" in an appropriate system of beliefs. All such views, applied to self-knowledge, are subject to the regress outlined in the text, given internalist assumptions. (I am indebted here to Crispin Wright.)
8. I am inclined to believe that (at least part of) what is going on in the famous passage that concludes Wittgenstein's discussion of rule-following in his *Philosophical Investigations* is an argument to this effect. The passage reads:

   > It can be seen that there is a misunderstanding here from the mere fact that in the course of our argument we give one interpretation after another; as if each one contented us at least for a moment, until we thought of yet another standing behind it. What this shows is that there is a way of grasping a rule that is *not* an *interpretation* . . . (*PI*:201)

   The textual evidence strongly indicates that Wittgenstein uses the term "interpretation" to mean "hypothesis as the meaning of." Read this way, the passage says that the moral of the rule-following "paradox" is that there must be a way of grasping the content of a mental event without having to form hypotheses as to its content.

   This is not the occasion to say what else might be going on in that passage or to defend this reading in greater detail.
9. This point is made in Tyler Burge, "Individualism and Self-Knowledge," *Journal of Philosophy,* 85.1 (Nov. 1988): 654–655.
10. Burge and Davidson are explicit about this.
11. The first version of the slow switching story involves questions that admit of no easy answer. Suppose both earthian and twin-earthian thoughts are simultaneously available to you. And suppose you think a thought that you would express with the words "I have arthritis." How is it determined whether this particular thought token involves the concept *arthritis* or the concept *tharthritis*? (This is not a question about how you would *know* whether it involved the one or the other; it's a question about what *makes it true* that it involves the one and not the other.) There seems to be no simple answer. It certainly does not seem right to say, for reasons that underlie the intuition that quick switching wouldn't suffice for change of content, that it is simply a function of the environment in which the thought is tokened. Nor are there other obvious dimensions of difference to appeal to: *ex hypothesi,* thoughts with the different contents would have

exactly the same functional roles, the same linguistic expression and the same associated qualitative episodes (if any).

12. It is no objection to this argument to point out that, on *this* way of telling the switching story, S cannot even frame the hypothesis he is called upon to exclude. Someone may not have the concept of counterfeit money, but if there is a lot of counterfeit money in his vicinity, then he must be able to exclude the hypothesis that the coin in his hand is counterfeit before he can be said to know that it is a dime. The fact that he cannot so much as frame the relevant hypothesis does not absolve him of this requirement. In any case, any residual worries on this score can be averted, if necessary, by telling the switching story in the alternative way outlined in the text.
13. See, for example, Andrew Woodfield's remarks in the "Introduction" to his collection *Thought and Object* (Oxford: Clarendon Press, 1982), viii. See also: Anthony Brueckner, "Brains in a Vat," *Journal of Philosophy,* 83. 3 (March 1986): 148–167.
14. See Tyler Burge, *op. cit.*, p. 651. See also his "Cartesian Error and the Objectivity of Perception," in *Contents of Thought,* Robert Grimm and Daniel Merrill, eds. (Tucson: University of Arizona Press, 1988).
15. This would be false if a language of thought hypothesis were true.
16. Tyler Burge, "Individualism and Self-Knowledge," p. 660.
17. D. Davidson, "Knowing One's Own Mind," *Proceedings of the American Philosophical Association* (1986): 453.
18. Jealousy is being used here merely for illustrative purposes. For reasons that are touched upon briefly below (see note 23), I actually rather doubt that judgments about jealousy are self-verifying in the sense bruited in the text.
19. Wittgenstein remarked, famously, that:

    > It cannot be said of me at all (except perhaps as a joke) that *I know* I am in pain. What is it supposed to mean—except perhaps that I *am* in pain? (*PI* 246)

    The remark has struck most philosophers as extremely implausible. A truth it may harbor, however, is that if (as seems right) it is constitutive of being in pain that one know that one is, then knowing that one is cannot count as a cognitive achievement; one doesn't count as being in pain unless one knows it. As against Wittgenstein, I am not sure that this point is best captured by denying that judgments about pain constitute 'knowledge'.
20. This observation is made in D. H. Mellor, "Conscious Belief," *Proceedings of the Aristotelian Society* (1978).
21. See Tyler Burge, "Individualism and Self-Knowledge"; Donald Davidson, "Knowing One's Own Mind"; and John Heil, "Privileged Access" in *Mind* 42.386, (April 1988): 238–251.

    An interesting proposal, that seems to me to fall somewhere in between a substantial and an insubstantial conception as defined here, is outlined by Crispin Wright in his "Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics," in *Reflections on Chomsky,* A. George, ed. (Oxford: Blackwell, 1989), 233–264. The proposal deserves extensive separate treatment. For a brief discussion see my "The Rule-Following Considerations" forthcoming in *Mind* (October, 1989).
22. Burge, *op. cit.*, p. 660.
23. This explains why second-order judgments about sudden wants or momentary frights cannot be self-verifying: these events are not mental performatives in the required sense. They cannot be brought about by the mere thinking of a second-order judgment; hence, they cannot be incorporated into a second-order judgment in the way required for self-verification.
24. *op. cit.*, p. 659.
25. Obviously, this barely scratches the surface of the various issues that crop up here. A proper discussion would include, among other things, an account of what Burge's self-

verifying judgments *do* constitute, if not a species of knowledge. Limitations of space prevent me from taking matters further in this paper.

26. For valuable comments on an earlier draft, or for helpful discussion of the issues, I am very grateful to David Velleman, Stephen Yablo, Barry Loewer, Jerry Fodor, Jennifer Church, and Crispin Wright.