10 Mental Content and Hot Self-Knowledge

Bernard W. Kobes

In this essay I explore the implications of a view of conscious thought as an activity that a thinker *S* engages in, rather than as something that *S* undergoes or as something that merely occurs in *S*. Such an account of cognitive agency promotes our understanding of a thinker's *authority* in making certain attributions of mental states and events to himself. The theme will be that a thinker's authority stems from two facts: first, that his self-attributions are not merely passive registrations of inner goings-on but have instead a certain performative character, and second, that he will typically have a kind of spontaneous knowledge of his performances.

In some of the recent philosophical literature, authoritative self-attributions have been seen as problematic in view of externalist accounts of mental content. I shall argue that the view of conscious thought as active, as having a certain performative character, can deepen our understanding of how it is that mental content is externally fixed. For example, it helps us respond to certain puzzles about how to ascribe mental content in Twin Earth scenarios, puzzles that call into question whether we have genuine knowledge of our current, conscious thoughts.

Assumptions and Background

I will assume that we can usefully distinguish mental *content* from mental *relation*. Thus, in the linguistic schema for ascribing propositional attitudes, ⌜*S* φs that *p*⌝, the mental content is given by the sentential complement *p*, while the mental relation is given by the predicate φ—believes, desires, hopes, wishes, doubts, and so on. The mental relation is broadly functional in nature. A mental relation is a characteristic kind of role that the content, or rather the *thinking* or *representing* of that content, plays in the thinker's cognitive economy. If, for example, a thinker on reflection moves from doubt to belief with respect to some content *p*, then one and the same content representation plays first one broadly functional role, then another.[1]

Mental relations may be broadly categorized into two kinds, depending on their *direction of fit*. Speech acts have been described as having either world-to-word or word-to-world direction of fit. Applying the distinction to Elizabeth Anscombe's vivid example, we may say that a list of grocery items prepared by a man going shopping has world-to-word direction of fit, while the same list of items prepared by a detective who follows him around in the grocery store has word-to-world direction of fit (Anscombe 1957, p. 56). (The direction of fit is typically the opposite of the direction of causal flow.) The distinction may also be applied to mental relations. Some, such as belief and appearance, have

mind-to-world direction of fit, while others, such as intention and desire, have *world-to-mind* direction of fit.

A mental relation's direction of fit is a matter of how we conceive the "responsibility" for success of a mental state or event involving that relation. In the case of belief and appearance, we hold fixed the state of the world, and it is the "responsibility" of the belief or appearance, or the processes that produce them, to effect a match between the mental content and the world. So the direction for belief and appearance is *mind-to-(fit)-world.* In the case of intention and desire, we hold fixed the mental state, and it is the "responsibility" of the world, or our actions on it, to effect a match between the mental content and the world. So the direction for intention and desire is *world-to-(fit)-mind.* Following Lloyd Humberstone, I will call relations like belief and appearance *thetic*, and relations like intention and desire *telic* (Humberstone 1992).

I further assume that Putnam, Burge, and others have established that many of our thought contents are externally fixed. So it is possible to vary in imagination the thinker's thought contents—even his de dicto thought contents—simply by varying his physical or social environment, while holding fixed his entire history of physical states of body and brain, "raw" phenomenal states (if such there be), and "narrow" behavioral and functional states.[2] But I will remain neutral among a variety of more specific theories about how mental content may be externally fixed. So I will not try to adjudicate among, for example, causal or historical chain theories, teleological theories, or covariation theories of mental content.

Despite the popularity of externalist views of mental content, philosophers are by and large impressed by the immediacy and authority of our knowledge of our own current conscious mental contents. If *S* currently and consciously thinks "Snow is an unnecessary freezing of water," then *S* knows the intentional content of his thought immediately—that is, without the aid of inference or observation—and with an authority that no one else has with respect to *S*'s thoughts. There is a prima facie appearance of tension here. For if *S*'s thought contents are fixed partly by conditions external to *S*, would it not be incumbent on *S* to investigate those external conditions in order to know what his own thought contents are? And if self-knowledge requires investigation of external conditions, then *S*'s knowledge of his own thought contents is neither more immediate nor more authoritative than anyone else's knowledge of *S*'s thought contents (see Woodfield 1982, viii).

The by now familiar reconciling response to the tension between content externalism and first-person authority is, very briefly, that the very external conditions that help fix the contents of *S*'s first-order thoughts also, and simultaneously, help fix the contents of *S*'s thoughts about his thoughts. In cases of what Tyler Burge calls *basic self-knowledge*—as when I judge that I am thinking, with this very thought, that water is more common than mercury—I use the same content in the higher-order thought about my thought as I use

in the lower-order thought about the world. The higher-order thought is self-referential and self-verifying; thinking it ensures its own truth. The external conditions on my thought content affect the first-order thought and the higher-order thought equally. The thinker need not know in a discursive manner what these conditions are in order to have a knowledgeable higher-order thought about his first-order thought.

These observations successfully block, I think, the simplest line of thought from externalism to our not having immediate and authoritative knowledge of occurrent, conscious mental contents. Let me call this the *Burge—Heil reconciliation story*, after two philosophers who presented early versions of it.[3]

Doubts and Discontents

I want now to collect and briefly review from some recent literature a number of doubts and discontents about the Burge–Heil reconciliation story. I will then propose a revised and extended version of the account, with the goal of answering these doubts and discontents. Answering the doubts and discontents will put externalism in a somewhat broader framework in the philosophy of mind, which, though controversial, will prove to have significant theoretical attractions.

The first source of doubt and discontent that I will consider is inspired by an argument of Brian Loar (1994).[4] Consider *S*'s reflexive judgment, "I am now thinking that Socrates drank some hemlock." Suppose *S* takes as given the existence of Socrates and hemlock. Still, *S* can think that Socrates drank some hemlock only if he stands in certain externally determined causal or historical relations to those objects. This reflection may cause *S* to suspect that his knowledge of his own mental content, even given the existence of Socrates and hemlock, is merely empirical. For, *S* reasons, it is open to doubt on empirical grounds that he himself stands in the appropriate external causal or historical relation to Socrates and to hemlock. So the issue, as Loar sets it up, is how *S* could reassure himself of the apriority of an inference from the existence presuppositions, that Socrates and hemlock exist, to the reflexive judgment "I am now thinking that Socrates drank some hemlock."

Now let us suppose that *S* knows that such reflexive thoughts are always true, given the existence presuppositions, since the content of the first-order thought and (part of) the content of the higher-order thought must be identical. *S* knows the Burge–Heil reconciliation story. The inference from existence presuppositions to reflexive judgment is, we may say, pragmatically valid. That is, if a thinking of the premise is true in a certain context, then a thinking of the conclusion would be true in the same context, even though the propositions expressed by premise and conclusion are related only contingently. But how can *S* assure himself, by way of such considerations, that the inference does not depend in some subtle way on his knowledge of external causal or historical relations that he bears to

Socrates and to hemlock? What *S* knows is that all thoughts of a certain form—including the thought "I am now thinking that Socrates drank some hemlock"—are true. But it would seem that in order to use that information to resolve his empiricist doubt, *S* would already have to know that he is thinking a thought of that reflexive form! And *that* knowledge may, for all we have said, depend on empirical knowledge of external causal or historical relations.

The difficulty, then, is that *S* cannot use his knowledge of the Burge–Heil reconciliation story in a wholly a priori demonstration, given the existence assumptions, that he is thinking that Socrates drank some hemlock. Loar imagines the thinker reasoning as follows:

> If I am now thinking that I am now thinking that Socrates drank some hemlock, that reflexive thought must be true, and so it would be the case that I am now thinking that Socrates drank some hemlock. Indeed. But this merely pushes the question back a stage. . . . [Given the coherence of the doubt that I stand in the appropriate relations to Socrates and to hemlock, how] am I to decide without further external information whether I am now thinking that I am now thinking that Socrates drank some hemlock? (Loar 1994, p. 64)

Extrapolating from Loar's argument, it would appear that if the Burge-Heil reconciliation story is all *S* has to work with, then *S* would be launched on an infinite regress if he tried to construct a wholly a priori justification for his inference from the existence of Socrates and hemlock to the conclusion that he is thinking that Socrates drank some hemlock. To demonstrate to himself a priori that he is thinking that Socrates drank some hemlock, he would first have to know that he is thinking that he is thinking that Socrates drank some hemlock. And presumably, to demonstrate to himself a priori that he is thinking *this*, he would first have to know that he is thinking that he is thinking that he is thinking that Socrates drank some hemlock. And so on. This is an unacceptable result; for surely if *S* can make this inference a priori, and if he has the requisite reasoning ability, he should be able to exhibit the inference to himself as a priori without regress. Moreover, the same should be true if we substitute for 'a priori', 'immediate[ly] and authoritative[ly]'.[5]

A second source of doubt and discontent concerns the scope of the reconciliation story. Burge discusses only basic self-knowledge, as when I judge that I am explicitly thinking, with this very thought, that water is more common than mercury. But surely our immediate and authoritative self-knowledge goes beyond that provided in this paradigm. For one thing, it seems that we have self-knowledge of what we are thinking even when we are thinking ordinary conscious thoughts about the world. In such cases we are not thinking explicitly reflexive thoughts. Yet we know what we are thinking. The explicitly reflexive thought of basic self-knowledge may be such that, as Burge says, thinking it ensures its own truth. But how does this help to explain our knowledge of our current conscious

thought contents when we are *not* thinking conscious reflexive thoughts in the form of Burge's paradigm?

Paul Boghossian has pressed a variety of other doubts about the extendibility of the reconciliation story (Boghossian 1989). Boghossian argues that the account does not explain our knowledge of the mental relations of our own mental states. He notes that "[s]elf-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" (Boghossian 1989, p. 21). More generally, although a thinking of the first-order thought content is entailed by a thinking of the higher-order thought content, the relational component of the first-order thought is not entailed by the thinking of the higher-order thought. So this aspect of immediate and authoritative self-knowledge remains unexplained.

Other doubts about extendibility loom. It seems we have authoritative self-knowledge with respect to the contents of immediately past mental events. How could the Burge–Heil reconciliation story account for this? As Boghossian points out, a memory-based judgment about what I was thinking *just now* is not self-verifying. The thinking of the first-order content can be a part of the thinking of the second-order content only if the two contents are absolutely coincident (Boghossian 1989, p. 21).

Moreover, the Burge–Heil reconciliation story appears not to explain our authoritative knowledge of our standing mental states. If I judge that I *believe* that writing requires concentration, my judgment is authoritative. But the standing state of which I judge is extended in time, so it is difficult to see how my judgment could be self-verifying or how the first-order standing state could be a part of the second-order judgment (Boghossian 1989, pp. 21–22).

So the second source of doubt about the Burge–Heil reconciliation story is really a family of concerns about its extendibility beyond the paradigm of basic self-knowledge.

Boghossian also charges—and this is my third source of doubt—that the reconciliation story renders our self-knowledge "cognitively insubstantial" when in fact self-knowledge is a genuine cognitive achievement. Self-knowledge varies in quantity and quality depending on the effort, attention, and skill that *S* puts into it. But if we have self-knowledge simply in virtue of the self-verifying nature of Burge's paradigm of basic self-knowledge, in which the content of the first-order thought is a part of the content of the higher-order thought, then self-knowledge requires neither observation nor inference based on observation. Yet it concerns contingent matters of fact. In these respects self-knowledge would be like the knowledge expressed by the sentence 'I am here now'. How, then, can self-knowledge be subject to cultivation or neglect? Given the Burge–Heil reconciliation story,

self-knowledge becomes problematic because it is in fact so imperfect, not because it is so perfect (Boghossian 1989, pp. 17–19).

Finally, there are puzzles that arise from reflection on so-called slow-switching Twin Earth cases. On Twin Earth they have no water; instead of H_2O, they have some other, superficially similar liquid with chemical composition XYZ, which they refer to by the word-form 'water', but which we on Earth may call twater. Suppose *S* is surreptitiously transported from Earth to Twin Earth. After a sufficiently long period of acculturation—say three decades of verbal interaction with his new cohorts, drinking and bathing in the liquid they call 'water', and the like—*S* will have acquired the relevant Twin Earth concept, which of course he now also expresses with the word-form 'water'. Thinking back to an occasion of over three decades ago, *S* remembers thinking a thought that he then expressed using 'water'. Now *S* does not *know* what he thought back then. Yet he has forgotten nothing! How can this be? Boghossian suggests an answer: perhaps *S never* genuinely knew what he was thinking (Boghossian 1989, pp. 13–15, 22–23).[6] So reflection on slow-switching scenarios supplies a fourth source of doubt about the Burge–Heil reconciliation story.

The Burge–Heil story answers a certain kind of very direct challenge to the compatibility of content-externalism and first-person authority. But as a positive account of self-knowledge it is rather thin, and that makes it difficult to see how to extend the account beyond the paradigm of basic self-knowledge. My goal here is first to set the reconciliation story within an account of the thinker as cognitive agent, and then to argue that the result gives us resources to respond to the four sources of doubt and discontent.

Cognitive Agency

To some philosophers it has seemed that we are doxastic agents, that what we think is up to us in certain ways. Others have seen our mental lives as having a more passive character. This contrast is vividly manifested in two views of the relationship of the thinker to his beliefs. Descartes, for example, held that belief involves assent by the will. Descartes seems even to have held that the will is perfectly free to assent to any proposition it chooses, with the possible exception of those rare occasions when it perceives clearly and distinctly. He writes:

> The will may be termed infinite; for we never observe any possible object of another will (even the immeasurable will of God) that does not also fall within the range of our own will. (Descartes 1971, p. 187)

> The existence of freedom in our will, and our power in many cases to assent or dissent at our pleasure, is so clear that it must be counted among the first and most axiomatic of our innate notions. (Descartes 1971, p. 188)

For Hume, on the other hand, belief is a passive phenomenon. Hume writes:

> [B]elief consists merely in a certain feeling or sentiment; in something, that depends not on the will, but must arise from certain determinate causes and principles, of which we are not masters. (Hume 1888, p. 624)

Each view finds support in certain elements of a reflective first-person perspective. On Hume's behalf, it must be said that the phenomenology of belief powerfully suggests that I cannot believe, just like that, anything I want to. On the other hand, my beliefs don't feel like they simply "happen" to me. Rather, *the world* often feels like it simply happens to me, and this may be sufficient to explain my sense, to the extent that I have it, of being passive with respect to belief. In fact, when I deliberate about what is the case, I seem to construct my beliefs about the world. Often enough it is a project that engages me as a person.

So Descartes's view is plausible in at least these respects: we are doxastic agents, and in learning, and in changing and updating our beliefs, we act on our store of beliefs. Taking cognitive agency seriously pays theoretical dividends: it can help explain some otherwise puzzling features of our self-knowledge. For under the right conditions, the commitment involved in agency allows for a spontaneous justified belief on the part of the agent about what he is doing.

If *S*'s belief and desire together cause an action that they jointly make rational, then *S* need not, as a matter of conceptual necessity, be aware of the rational connection between the belief and desire, on the one hand, and the action, on the other. There is no requirement from the very nature of rational causation that *S* be aware of the rational connection that links his belief and desire to his action. However, *S* may *in fact* be aware of his belief and desire and the rational connection between them and his action, and this awareness of a rational connection may be efficacious in the production of the action. Let us use the term *higher-order rational causation* for such cases.

Even in higher-order rational causation, *S* need not be aware of the rational connection between his beliefs and desires, including now his second-order beliefs about first-order rational connections, on the one hand, and the action, on the other. The point here is again to warn against a view of rational mental causation as requiring *S* always to be aware of the rational connections between the causal relata, for this view seems to generate a vicious regress. At some level rational causation, even higher-order rational causation, is "tropistic," in that it ultimately involves causal processes not under the rational supervision of the agent.[7] Yet it may be an important fact that conscious agents engage in higher-order rational causation, even if this is not required by the very idea of the rational causation of action. Higher-order rational causation may be typical for us; it may be even be essential for *conscious* action.

Moreover, action may be turned inward; we can change not only the world but also our minds. Sydney Shoemaker has argued that the rational updating of beliefs and intentions in light of new experience requires that the thinker have mental access to the contents of his current beliefs and intentions.[8] And there may be advantages, from the standpoint of efficient design and engineering, for a rational device to have information about its mental states and events and their logical and evidential relations to one another.

It would of course be a mistake to portray our mental lives as always deliberative. Many, perhaps most of our thoughts simply occur to us without deliberation. But most of our ordinary bodily actions are likewise unaccompanied by deliberation. Bodily actions may be the products of habit, or alternatively, whimsy or caprice. In a choice between several equally attractive alternatives we are capable of making what seem to us random selections, and many of us vary our routines now and then in spontaneous ways. Yet, tics and involuntary tremors aside, all such bodily actions still seem up to us; we know what we are doing, and we maintain a sense of being their author.

The same is true of cognitive agency. Though a thought may simply occur *to* me, it does not simply occur *in* me. I know what I am thinking, and obsessive-compulsive and hallucinatory thoughts aside, I have a sense of being not merely the possessor of my thoughts but their author. A thought occurs to me: "Oh, there's Mark Richard in a red sweater." I am certainly not responsible for the fact that I thereby represent; I am not even, in the relevant sense, the author of the red visual sensation that triggers and accompanies my thought. But I am the author of the thought, and I know what its intentional content is because *I get to say* what it is. The naturalness of these locutions, even for a casual thought in passing, marks the sense of cognitive agency with which our explanation of authority begins.

Commitment and Spontaneous Beliefs

Stuart Hampshire argues in his book *Freedom of the Individual* that in having an intention to *A* an agent *S* often has a spontaneous belief that he will *A* (Hampshire 1975).[9] A spontaneous belief is a belief that is not derived from or causally based on anything else that the believer takes as a premise for the belief. If, for example, I am contemplating whether to buy some flowers on my way home from work, and then form the intention to do so, I have a spontaneous belief that I will buy flowers on my way home from work—a belief that is not derived from or causally based on anything else that I take as a premise for the belief. And such beliefs are often justified. Generally, in a case in which I believe with justification that *A*-ing is fully within my power, and in which I intend to *A*, and possess no special reason to doubt that I will carry out my intention, I have a spontaneous justified belief that I will *A*.

For me to have an intention to *A*, in a case in which I think *A*-ing is within my power, I have to think of it as up to me, as something I make happen. An intention is a commitment to act, and typically as soon as I am so committed I have the belief that I will *A*. If in having the putative intention to *A* I do not eo ipso believe that I'll *A*—in a case in which I think it fully within my power and in which I possess no special reason to doubt that I will carry out the intention—then I am not exercising my commitment in the way that is paradigmatic for a genuine intention. So I can and typically do believe that I will *A* just in virtue of exercising my commitment to *A*-ing, and not in virtue of a belief about the existence of my intention to *A* together with a general belief about my likelihood of success in intentions of this kind.

Granted, I might also reason that I will *A* based on my knowledge of the existence of my intention, together with the generalization that I always or almost always fulfill such intentions. I am typically in a position to step back from my practical commitment, acknowledge it thetically, and argue from the existence of the commitment, together with a general premise about how frequently I follow through on such intentions, to the conclusion that I will *A*. Such arguments, however, only bolster my justification for believing something that I may already know by a different channel. For if I argue from the existence of my intention, together with a general premise about my record of following through on intentions of this kind, then my intention must exist prior to and independently of my carrying out this train of thought. Moreover, my auxiliary premise about my record of following through on intentions of "this kind" would be sensitive to my exercise of the intention itself, the flexing of my resolve. And frequently my belief that I will *A* is coeval with the sheer exercise of my intention, prior to my knowledge of its existence.

The proximal rational cause of my belief that I will *A* is then a telic state, in Humberstone's terminology, a state with world-to-mind direction of fit, rather than a thetic state, a state with mind-to-world direction of fit. Let us call the first kind of knowledge that I will *A*, resulting proximally from the exercise of my commitment to *A*-ing, *hot self-knowledge*, and the second kind of knowledge that I will *A*, resulting from knowledge of the existence of my intention to *A*, together with general premises about my likelihood of success in intentions of this kind, *cold self-knowledge*.

In the fullness of time, if all goes well, an intention to act is translated into an intentional act, and the belief that I will *A* becomes the belief that I am *A*-ing. The earlier belief that I will *A* is dynamically continuous with my current belief that I am *A*-ing. If the earlier belief was hot, then typically so is my current belief. That is to say, my current belief that I am *A*-ing has among its proximal rational causes a telic state, a state with world-to-mind direction of fit, namely, my intention to now *A*, or to now be *A*-ing. My belief that I am *A*-ing need not and typically does not derive from a belief about the existence of my current intention together with a belief about its succeeding.

Certainly my belief that I am *A*-ing depends on complex empirical feedback. My belief that I am now buying flowers on my way home from work depends on perceptual feedback on my progress: the sight of the flower shop, the proprioceptive sense of walking into the store, the smell of the flowers, the sounds of the verbal transaction, the heft of the bouquet. Nevertheless, my belief that I am buying flowers remains a hot belief. The perceptual feedback serves a crucial auxiliary function. It enables me to check on my progress. But my buying flowers is a larger event than any of these perceptual atoms by itself, or the particular state or event thus perceived, or even the sum of perceived states or events up to any given time prior to the completion of the project. How do I know that I am buying flowers? Not, I claim, by an inductive extrapolation from the sum of the perceived states or events up to now, nor even typically from perceptual feedback up to now plus knowledge of my intention plus a general premise about my likelihood of success, but rather from perceptual feedback up to now plus the very exercise of my intention. Since its proximal rational causes include a state with world-to-mind direction of fit, my belief that I am now buying flowers is hot.

The view then is that hot belief of what I am doing or will do precedes any cold belief I might have about what I am doing or will do. The hot belief is prima facie warranted in the absence of compelling reason to consider the matter coldly. Moreover, it is plausible that my epistemic entitlement to the hot belief does not depend on any disposition I might have to justify it coldly. Indeed, a particular believer might not have the mental capacity coldly to justify his belief that he will *A*. It is true that if I have reason to think or ought to think that I often change or abandon intentions of this kind, or that circumstances are likely to arise in this case that will cause me to abandon or fail in my intention, then I cannot be said to be justified in my belief that I will *A*. If my hot belief could not withstand cold scrutiny then it is unjustified. But these are only "defeating" conditions. The relevant necessary condition for my hot belief to be justified is simply the absence of such defeating conditions, not any disposition of mine to check or know that they are absent.

Now in belief formation, I am passive with respect to those external facts that are beyond my control, yet it is a project of mine—it is "up to me"—to form beliefs that are faithful to them. Something roughly analogous is plausible about the linguistically expressible concepts that I exercise in thought. It is not up to me which concepts my sociolinguistic environment makes available to me, but given that a certain stock is made available, it is up to me which of those I employ in a given thought. The intuitive point is not so much that, when I exercise a concept, it's *my* concept, but rather that it is *my exercise* of some concept or other, and that I get to say which. This supplies a rough sense in which it is up to me what the content of a given thought is, and interpretation is accordingly constrained.

It is possible to explicate this sense in which it is "up to me" which concept I am exercising. Think of the mental relation to the content of self-knowledge—⌜I think that *p*⌝—

first of all as telic. My mental relation to the content ⌈I think that *p*⌉ has world-to-mind direction of fit—though now, of course, the relevant "world" is itself my mind, or some part or aspect of my mind. The content of self-knowledge is something that I thereby *make true*. My bearing a telic relation to the content ⌈I think that *p*⌉ explains my thinking that *p*.

I have argued elsewhere that, in the case of the belief relation, the representation in the higher-order thought of *p* as something believed often explains or accounts for the representation-of-*p*'s playing the functional role of belief (Kobes 1995). Similarly, I now want to suggest, at least in the case of linguistically expressible thoughts, a first-order thought's playing the role of a constituent in a self-regarding higher-order thought to which *S* is telically related helps explain or account for its having the first-order content that it has. For in virtue of *S*'s telic relation to the content ⌈I think that *p*⌉, *S* implicitly forges links between his current thought and thoughts that *S* attributes or could attribute to others, and thoughts that *S* remembers or could remember thinking.

These links are broadly syntactic in nature. They do not take the form of a revisable hypothesis, as in thinking that the words 'heather' and 'gorse' express or refer to the same property. Rather, they supply a formal constraint on the interpretation of *S*'s concepts, via a link resembling the syntactic link of two word-tokens that are counted as belonging to the same type. In this way *S* implicitly makes it the case that he is tokening a thought of a certain intentional type. For facts about *S*'s dispositions, given his telic relation to the content ⌈I think that p⌉, to make such links to the thoughts of others and to his own past thoughts, are among the facts in virtue of which his current thought has the intentional content that it has.

A thinker's ability to think a given thought content will in many cases depend on his ability to use a natural language. Even a thinker who cannot access the relevant word-forms, owing to forgetfulness or to aphasia, may think thoughts that he would not have been able to think had he not once used a natural language. Externalist accounts of intentional mental content, such as Burge's, key on such linguistically mediated thoughts. The truth conditions of *S*'s natural-language sentences help to fix the truth conditions of his thoughts, even if *S*'s mastery of those truth conditions is flawed or incomplete. An account of how the public intentional content of a natural-language sentence can fix the intentional content of a thought must avoid the trap of making the relevant thought or mental act metalinguistic; it is not the thought or the bet that a certain sentence is true.

I suggest that we can deepen our understanding of these matters by treating *S*'s relation to a self-attribution as telic. *S* participates in public practices of assertion and attribution of belief. Another person's guileless assertion is taken in this practice to be a sufficient criterion for an attribution of belief to him. Such attributions of beliefs to others have publicly available truth conditions in virtue of being couched in a common natural language.

Each is in a position to attribute to himself the very thought that he can attribute to others as part of the public practice. *S* is capable of thinking, ⌜*R* thinks that *p*⌝; plugging himself into the subject position, he can think ⌜I think that *p*⌝. *S* bears a telic relation to that self-attributive content; it has a performative character. *S* thereby makes it the case that he thinks that *p*.

P. F. Strawson and Gareth Evans have been concerned to argue that the occurrence of the predicate in a self-attribution must be seen as univocal with its occurrence in a corresponding third-person attribution.[10] For example, Evans (1982) has argued, by way of his "Generality Constraint," that the self-attributing thinker must be capable of understanding his attribution as having been made to a third person. Even given this point, it might not be wholly clear why we should not individuate first-order thoughts individualistically and treat all attributions as nonindividualistically typed but the best we have for purposes at hand, or as serving some purpose that cuts across psychological explanation, for example, the transmittal of information.[11] On the conception I have been advocating, mental content is "inherited" from the self-attribution to the first-order psychological state. By committing himself to a mental content, *S* applies to himself a public practice of content attribution that he already engages in with respect to others. The self-attribution has a performative character in virtue of *S*'s telic relation to its content. Thus the public truth-conditions of the attribution get a purchase on the individuation of his own psychological state.

Some aspects of a linguistically expressible concept's causal role help fix it as belonging to a certain intentional type, whereas other aspects of its causal role derive from *S*'s ability to use the concept properly. (I set aside issues about error and misuse.) A concept has a causal role only in contexts of full thoughts. The current proposal is, in effect, that certain aspects of a concept's causal role in virtue of the telic higher-order thought—dispositional links to attributions to others and to one's past self—establish or at least constrain its intentional type. The dispositional links help to fix the concept's intentional type, and underwrite interpersonal comparability of concepts, and intrapersonal comparability of concepts over time. Intentional type in turn fixes the proper causal roles of first-order thoughts and concepts—and derivatively, their actual roles, insofar as the thinker is able to use his concepts properly.

Moreover, by an internal analogue of Hampshire's thesis about spontaneous knowledge, *S* knows that his thought has the content *p*, in the way that one knows things one intends to do or is intentionally doing. The telic relation *S* bears to the content ⌜I think that p⌝ is a form of commitment, and in being so committed *S* acquires a spontaneous belief with the same content, namely, that he does think that *p*. His self-knowledge about the content of his belief is nonobservational and noninferential, and in particular does not depend on his knowing about his own telic higher-order thought. Thus an analogue of Hampshire's

thesis gives *S* authoritative knowledge about that over which *S* has a certain kind of control. Authoritative self-knowledge of the relevant kind is the product of the performative or telic character of the higher-order thought, together with an internal analogue of Hampshire-style spontaneous belief.

Authoritative self-knowledge derives, then, not from *S*'s being well-positioned to observe his own thoughts, and even less from a polite convention of attribution, but from *S*'s being a cognitive agent, together with the spontaneous knowledge that characteristically accompanies agency. Authoritative self-knowledge in this sense may be only one species of immediate or privileged self-knowledge. There may be kinds of immediate or privileged self-knowledge—for example, of *S*'s sensations—that are not in the same sense authoritative. For they do not derive from the authority invested in *S* as a cognitive agent, but only from *S*'s immediacy or privilege in being the one in whom the sensation is "broadcast."

Smoldering Self-knowledge

Let us return to one of Boghossian's challenges to the extendibility of the Burge–Heil reconciliation story. Boghossian notes that in Burge's paradigm of basic self-knowledge the higher-order thought must be absolutely coincident with the thought it is about. But, he argues, this is a very special condition; we also have direct knowledge about our immediately past thoughts, and Burge's account does nothing to explain how that is possible. If, for example, at t_1 I think that writing requires concentration, then at t_2, a very short time later, I can with authority judge that I just now thought that writing requires concentration. But my judgment at t_2 is not self-verifying; it is conceivable that I make the t_2 judgment without having made the t_1 judgment. The first-order thought at t_1 is not a part of the second-order thought at t_2. Yet I know what I just now thought noninferentially and authoritatively. So, it appears, Burge's proposal is incapable of explaining some paradigm cases of direct self-knowledge (Boghossian 1989, pp. 21–22).

Let us assume that self-knowledge of past thoughts can be direct, in the sense that it need not and often does not rely on any kind of observation of a memory trace, or inference based on observation. This assumption of course sharpens Boghossian's challenge, since his point is precisely that the Burge–Heil reconciliation story cannot explain the directness and authority of some of our knowledge of our own past thoughts. Often the thinker simply makes a memory-based judgment of what he thought without making his memory or memory-trace the subject of any observation or inference, and he is epistemically justified in doing so.

But what is the intentional content of the memory? It is not simply the content of the first-order thought at t_1; that is, it is not simply: Writing requires concentration. For that

would constitute remembering a general fact, that *writing requires concentration*, whereas what the thinker remembers is a specific event, namely, that *I just now thought that writing requires concentration*. So the memory trace that was formed at t_1 and comes to consciousness at t_2 has the content of a higher-order thought.

But how did a memory trace with this content come about? My suggestion is that it is the *smoldering* remains of a hot self-belief at t_1. It came about at t_1 as a spontaneous concomitant of the act of judging that writing requires concentration. It was stored in the way that memories are stored, and came to consciousness at t_2. (Or perhaps it was conscious all along, active in "working memory"; Boghossian's case does not specify this one way or the other.) So although self-knowledge at t_2 is not, strictly, hot self-knowledge, it nevertheless depends on prior hot self-knowledge, preserved in memory.

We do typically remember ordinary conscious but non-self-conscious thoughts. We not only remember their first-order intentional contents, but we remember thinking them. In Burge's paradigm of basic self-knowledge, as when I judge that I am now thinking, with this very thought, that water is more common than mercury, the higher-order thought is presumably *itself* a conscious thought. But one theoretically attractive way to extend Burge's paradigm of basic self-knowledge to conscious but non-self-conscious thought is to suppose that a necessary condition for a mental state or event m's being conscious is that it be accompanied by a perhaps unconscious occurrent thought to the effect that one is in that very state m.[12] When a bit later one remembers that one just now thought that p, that is the coming to consciousness of a formerly unconscious higher-order thought. Thus we would have an explanation of the apparent immediacy of some of our knowledge of past conscious but non-self-conscious thoughts.

The idea of smoldering self-knowledge illustrates how the paradigm of basic self-knowledge may be extended to account for the authority, such as it is, of our knowledge of our own past thoughts. Memory at t_2 of past thoughts is in a derivative sense authoritative, for it depends on the authority of S's hot self-knowledge at t_1. At the same time, all memory, including memory of past mental events, is a fallible process, and subject to cultivation and neglect. (Trying to recall the thought that caused me to be vaguely depressed a few moments ago, I may think it was the thought of my friend's misfortune, when in fact it was the thought that a similar misfortune might befall me.) Even when I identify a remembered thought not by its causal relations but by its content, I may simply misremember that content. But when I do correctly remember the content (I recall, "Yesterday I thought: hypochondria is the only disease I haven't got"), my current thought, being the trace of earlier hot self-knowledge (the day-old trace, that is, of yesterday's thought, "I think: hypochondria is the only disease I haven't got"), partakes of its authority.

More generally, note that there is an uncomfortable tension between, on the one hand, Boghossian's objection that the Burge–Heil reconciliation story makes self-knowledge too

easy, hence cognitively insubstantial, and, on the other hand, his charge that the account does not explain the range and variety of direct, authoritative self-knowledge. The former objection depends on our self-knowledge being error-prone and fallible, while the latter objection depends on our self-knowledge being direct and authoritative in a wide range of cases. The general strategy of my response exploits this tension and is illustrated by the idea of smoldering self-knowledge. Though I will not carry out the strategy in full detail here, I will outline it briefly.

Basic self-knowledge is the central phenomenon that a variety of extended kinds of self-knowledge exploit. Basic self-knowledge is indeed cognitively insubstantial in several senses: it comes easily and errors are difficult or impossible to imagine; it does not depend on observation or inference based on observation; and it is not sensitive to the degree of the subject's attention, training, self-honesty, or skill. The case of conscious but non-self-conscious occurrent thought can be subsumed under the same rubric by postulating unconscious occurrent higher-order thoughts. The three extensions that Boghossian discusses—knowledge of the mental relation, knowledge of past mental states, and knowledge of standing mental states—build on a foundation of basic self-knowledge but extend it in certain ways. The extensions introduce occasions for error, and occasions for varying degrees and quality of self-knowledge, depending on attention, training, self-honesty, and skill. So extended self-knowledge, unlike basic self-knowledge, is cognitively substantial in several senses, while deriving a partial authority from the basic case.

"Slow-Switching" Twin Earth Puzzles

Bertrand Russell once suggested that a good test of a philosophical theory is how well it handles puzzle cases. In that spirit, let us suppose that at t_1 *S*, an ordinary Earthling, thinks "I am thinking that water is a liquid." His self-knowledge is direct and authoritative. Now suppose that he is surreptitiously switched to Twin Earth. He interacts with the speakers there, and with the twater there. Eventually he acquires the concept that Twin Earthlings express by the word 'water', that is, the concept twater. But now, at t_2, he can still think back to t_1. Although he has forgotten nothing, he no longer knows what he thought at t_1. Boghossian writes:

> No self-verifying judgment concerning his thought at t_1 will be available to him then. Nor, it is perfectly clear, can he know by any other non-inferential means. . . . 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 t_1, and if at (some later time) t_2, *S* remembers everything *S* knew at t_1, then *S* knows that *p* at t_2. 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. . . . 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. (Boghossian 1989, p. 23)

Note the purely hypothetical role of the Twin Earth switching scenario in this argument. The argument exploits only the bare logical possibility of Twin Earth switching. If the argument is a good one, it shows that Burge's self-verifying judgements do not constitute genuine knowledge even if we are justifiably certain that no Twin Earth switching actually occurs.

First, let us consider the puzzle in the following version: we stipulate that *S* knows at t_2 that he has at some past time been subject to Twin Earth switching. Call this *knowledgeable slow switching*, since *S* knows that he has been switched. (Actually the case depends only on *S*'s believing or suspecting that he has been slow-switched, or doubting that he has not been slow-switched.) But *S* simply does not know if he is now on the same planet that he was on at t_1 or not. In particular, he does not know whether he was switched before or after t_1. He uses both 'water' and 'twater'. But he remembers the t_1 event, and in fact he wonders, "Was I thinking at t_1 about water or twater?" This seems to be the version of the puzzle that Boghossian has in mind.

But *S* does have available to him the same concept that was available to him at t_1. He is perfectly free to exercise the same concept at t_2. At t_2, thinking back to t_1, he may invent a new word for the substance he remembers so well, say, 'mwater'. He knows that mwater is either water or twater. He does not know which, but this is no defect in the concept *mwater*. In fact, there are a great many things he knows about mwater. For example, he knows that around the time of t_1 it often rained mwater. And he knows that mwater is a clear colorless liquid that quenches thirst. Are we to deny him this knowledge just because he does not know how to re-express it in terms of the words 'water' or 'twater'? That would be highly uncharitable; it seems to be perfectly good knowledge of the external world. And he knows one more thing: he knows that at t_1 he thought that mwater was a liquid. He knows this because he has a memory trace that results from the hot self-knowledge he had at t_1. This too is perfectly good smoldering self-knowledge. Even at t_1, it may interact with other things he believes about the world in inference and practical reason. Though he does not know how to re-express this bit of self-knowledge in terms of 'water' or 'twater', it should nevertheless be reckoned among his cognitive achievements.

Let us suppose, then, that in Boghossian's puzzle *S* never finds out that he was switched, and does not even suspect that he might have been switched. Call this, in contrast with the first case, *ignorant slow switching*. At t_1 *S* had the concept *water*; at t_2, he has the concept *twater*. At t_2, *S* thinks a thought that he would express as, "At t_1, I thought that water is a liquid." It seems to us that he does not really know at t_2 what he thought at t_1.

Yet he has forgotten nothing. Does Boghossian's argument, concluding that his thought at t_1 did not constitute genuine knowledge, go through now?

No. For one thing, Boghossian's "platitude" is plainly false. According to the platitude, quoted above, if S knows that p at t_1, and if at (some later time) t_2 S remembers everything S knew at t_1, then S *knows* that p at t_2. But S might acquire a new belief q that causes him to lose his old knowledge that p, even though S forgets nothing. Suppose that q constitutes evidence for the proposition $\neg p$, evidence that is misleading, as it happens, since p is true. Sherlock Holmes might have sufficiently strong evidence to *know* that the countess murdered the earl with a knife. Then he comes to believe that the countess was on the train to London at the time the earl was stabbed. Now Holmes no longer knows that the countess is the murderer, but he has forgotten nothing.

Note that the counterexample is compatible with the new belief q's being itself true. Even if q is true, its misleading evidential bearing on $\neg p$ can undermine the status of S's belief that p as knowledge. The countess was indeed on the train. Holmes's coming to learn this undermines his justification for his belief that the countess is the murderer. Moreover, the counterexample is compatible with S's continuing to believe that p. Holmes irresponsibly persists in his belief that the countess committed the murder. Luckily for Holmes, he is right: the countess had set up a fiendish knife-throwing contraption hooked up to a timing device. But Holmes's belief has lost its status as knowledge.

So Boghossian's "platitude" is quite false. In slow switching, we are not forced to choose between S's having forgotten something he once knew and S's never having had genuine knowledge. The lesson of our reflections on Holmes is that, if S no longer knows the content of what he once thought, this *may* be because some intervening event, such as ignorant slow switching, has destroyed the status as knowledge of some once-genuine knowledge that is still remembered.

But there is more to be said. At t_2, S has a self-belief that he could express by saying, "At t_1, I thought that water is a liquid." Does the thought S would express in this way constitute memory-based self-knowledge? On the one hand, we are tempted to say that it does, on the grounds that it results from a veridical memory trace of earlier self-knowledge. On the other hand, we are tempted to say that it does not, on the grounds that at t_2 S is exercising the concept *twater*, a concept he did not possess at t_1.

The answer depends, I think, on recognizing that S's thought at t_2, which he would express by saying "At t_1, I thought that water is a liquid," is equivocal. Unbeknownst to S, his thought at t_2 has two propositional contents simultaneously. The thought at t_2 is a result of S's cognitive agency at t_2 and therefore cannot escape including an exercise of the Twin Earth concept *twater*. At the same time, S's exercise at t_2 of the concept he expresses by 'water' was triggered by a memory event that is rooted in the Earth concept *water*. S's concept has historical roots in two speech communities, and in distinct kinds

of clear colorless drinkable liquid on two planets. So it is an equivocal concept, no matter how much twater has flowed under the bridge.[13]

We may represent *S*'s total corpus of beliefs at t_2 by way of a fragmentation strategy (cf. Lewis 1982). *S*'s total belief corpus at t_2 is broken into two large overlapping fragments. It appears to *S* as only one coherent belief corpus, because the concepts in terms of which he represents the corpus to himself include the same equivocal concept as the (first-order) corpus itself. If *S* had *n* concepts that were independently two-ways equivocal, perhaps through a series of *n* independent surreptitious switches to different linguistic environments, then *S*'s belief corpus would be broken into 2^n distinct fragments.

When a belief corpus is fragmented in this way, we may say that *S* has *weak* self-knowledge (i.e., self-knowledge in the weak sense) if he has self-knowledge according to at least one fragment, and that *S* has *strong* self-knowledge if he has self-knowledge according to all fragments. Then in the case at hand we can say that *S* has weak self-knowledge at t_2 about his belief at t_1. For on one fragment he has a belief that at t_1 he thought that water is a liquid, which is true, but on another fragment he has a belief that at t_1 he thought that twater is a liquid, which is false. In this way we can account for and honor two intuitions about Boghossian's puzzle case: that *S* holds an erroneous belief at t_2 about his thought at t_1, and that he in some sense *still knows* what he thought at t_1. But the conclusion that *S* had at t_1 no genuine self-knowledge is undermined.

Note that on this account, *S* has strong self-knowledge at t_2 about his thought at t_2, self-knowledge that he might express by saying, "I believe that I judge that at t_1, I thought water is a liquid." For on one fragment he believes that he judges that at t_1 he thought that water is a liquid, which is true; and on another fragment he believes that he judges that at t_1 he thought that twater is a liquid, which is also true.

More generally, the exercise of an equivocal concept is no barrier to strong, hot self-knowledge. Some time after a surreptitious switch from Earth to Twin Earth, given acculturation, *S* will have both the concept *water* and the concept *twater*, and will unwittingly exercise both simultaneously whenever he thinks a thought that he would express using the word 'water'. A cognitive agent in this situation has strong, hot self-knowledge. Of course *S* fails to know that his thought is equivocal, and this is a significant limitation on his discursive self-knowledge. But this is just another instance of the general point that a thinker can lack knowledge of the individuating conditions of his thought, even while those individuating conditions help to fix the identities of both his first- and second-order thoughts.[14]

My account of the two versions—"knowledgeable" and "ignorant"—of the slow-switching scenarios illustrates and supports the theory presented earlier of authoritative self-knowledge of mental content. According to that theory, the higher-order thought involves a telic relation to its content, and works by implicitly establishing links between

the self-attribution and attributions to others and to one's past self. Now in the case of knowledgeable slow switching, *S* establishes a link between his current t_2 thought (about his past t_1 thought) and the remembered t_1 higher-order thought itself. The concept, which *S* labels 'mwater', has the same intentional content as our concept *water*, and (unbeknownst to *S*) it is true of all and only instances of H_2O. Moreover, *S* takes his currently exercised concept to have its intentional type entirely fixed by that link, and *S* excludes links to the concept that he ordinarily, these days, expresses by the word-form 'water'. And it is precisely because of the performative character of *S*'s higher-order thought, in virtue of his telic relation to its content, that he can exercise a concept, like *mwater*, thus selectively linked.

On the other hand, in the case of ignorant slow switching, *S* establishes links between his current t_2 thought (about his past t_1 thought) and *both* the remembered t_1 higher-order thought itself, and other current t_2 thoughts. Because he has the authority to do so, his current thought becomes equivocal; it has both contents simultaneously. It is precisely because of the performative character of *S*'s self-attribution, in virtue of his telic relation to its content, that we are in no position to discount one set of links as mistaken. As reasonable interpreters of *S* we must therefore acknowledge both, and ascribe to him an equivocal concept. The performative character of self-attributions comes through with unusual clarity, I venture to suggest, in the unusual contexts of slow switching scenarios.

Boghossian imagines a slow-switching scenario in which Peter, hiking in northern New Zealand, encounters the tenor Luciano Pavarotti floating on the pristine waters of Lake Taupo. They chat briefly. Some years later Peter is switched to Twin Earth, where the word-form 'water' refers to XYZ, and 'Pavarotti' and 'Lake Taupo' refer to the twins of Pavarotti and Lake Taupo respectively. Peter is unaware of the switch, so this is a case of what I have called "ignorant" slow switching. Over decades he is insensibly acculturated to Twin Earth words and concepts, but his memorable encounter with Pavarotti stays with him. He reads and understands newspaper reports using the name 'Pavarotti' [which are about Twin Pavarotti], but his vivid and accurate memories of the long-ago encounter are about Pavarotti floating on water (Boghossian 1994, pp. 36–39; 44–45).

It would seem, then, that some of Peter's tokens of 'Pavarotti' refer to Pavarotti, while others refer to Twin Pavarotti, and some of his tokens of 'water' refer to H_2O, while others refer to XYZ. Yet Peter himself is unaware of this, and could become aware of it only by empirical investigation, and not by any a priori reflection on his mental life no matter how rational and thorough. Or so Boghossian interprets the case. The case shows, he claims, that externalist accounts of mental content conflict with a highly plausible principle of "transparency of difference," namely, that if two of a thinker's thoughts possess distinct intentional contents, then the thinker must be in a position to know a priori that they do. Worse, the case seems to show that externalism about mental content allows for the

possibility that two of a thinker's thoughts *of the same syntactic type* might have distinct contents, without the thinker being in a position to know a priori that they do.[15]

This would be an intolerable result for externalism. It allows, as Boghossian notes, the possibility of a perfectly rational thinker committing himself to logically invalid arguments. Reflecting on the properties of the stuff he calls 'water', Peter thinks, "Whoever floats on water, gets wet" (P1). As a general quantified thought by a naturalized Twin Earth resident, this true thought employs a concept referring to XYZ. Peter also thinks, "Pavarotti once floated on water" (P2). Based as it is on a vivid and veridical memory of his distant encounter with Pavarotti, this true thought employs a concept referring to H_2O. Peter concludes, "Pavarotti once got wet" (C). The argument is invalid, since the two premises equivocate on a key concept. But it will seem valid to Peter, and no amount a priori reflection could reveal to him that it is not.

Our account of how to attribute contents in cases of ignorant slow switching blocks this intolerable result in a plausible and externalistically acceptable fashion. Peter's thoughts, unbeknownst to him, have equivocal intentional contents. His thought "Whoever floats on water, gets wet" expresses two propositions at once, one containing the relational property *floats on twater*, the other the relational property *floats on water*. His thought "Pavarotti once floated on water" expresses two propositions at once, one involving Pavarotti, the other involving the tenor's twin. "The" argument is really two arguments, both valid (and probably both sound as well, if we may assume that Twin Pavarotti once floated on XYZ). Externalism is therefore not committed to Peter's being disposed to reason invalidly.

This interpretation of Peter as thinking equivocal thoughts is not an ad hoc defense of externalism. Although Peter's thought "Pavarotti once floated on water" is memory based, its subject concept is Peter's, and *he gets to say* whether he is exercising the same concept as he did when reading yesterday's newspaper report about the man it called 'Pavarotti' (i.e., our tenor's twin).[16] Our theory articulates this plausible and intuitive idea as follows: In making his memory-based judgment Peter is a cognitive agent, and therefore bears a telic relation to the content "I am thinking that Pavarotti once floated on water." This thought of Peter's derives its content at least partly from Peter's dispositions to link it to attributions that he might make to others, and to himself at other times, and not solely from the temporally distant man and liquid that are the causal provenances of his first-order memory trace. The telic higher-order thought has a performative character, and makes it the case that Peter's first-order thought has the content—or, in this case, contents—that it does. Given the phenomenon of spontaneous belief attendant on telic relations, Peter will know the contents of his thoughts, but this "strong" self-knowledge will also be equivocal in a way that masks from rational view the equivocation of his first-order thoughts.

There is, I think, no clear sense to be made of the idea that two of a thinker's (nonindexical, nondemonstrative) thoughts or concepts might be of the same syntactic type but have distinct contents. For the clearest notion we have of two thoughts or concepts being of the same syntactic type derives from the thinker's dispositions to link them in characteristic ways in higher-order thoughts in which they appear, and such links will constitute the thoughts or concepts as having the same intentional content.

One manifestation of this is a thinker's disposition to link concepts in deductive reasoning. Consider Peter's linking, in the argument above, the 'water' concept of (P1) with the 'water' concept of (P2). Intuitively, Peter's linking them in the argument manifests not merely his belief that he is exercising concepts with the same content, but his intention to do so. Of course, the same might be said about all equivocation in argument. One who thinks, "I have a duty to do what's right; I have a right to offer my frank opinion of your hat; therefore I have a duty to offer my frank opinion of your hat," in some sense intends the tokens of his concept *right* in the two premises to have the same intentional content. In many such cases the thinker can be brought to see his error by calling his attention to other examples that make his equivocation in this example salient to him. So a particular intention to exercise the same concept in two premises of an argument is not sufficient to constitute them either as belonging to the same syntactic type or as having the same intentional content. In Boghossian's case, however, where it is agreed that the syntactic type of the two 'water' tokens is the same, Peter's linking the 'water' concepts in his argument about Pavarotti is but one manifestation of a *complete* set of relevant dispositions. For the set of attributions to which Peter is disposed to link the higher-order thought corresponding to (P1), as containing a 'water' concept of the same type, is identical to the set of attributions to which he is disposed to link the higher-order thought corresponding to (P2). The performative characters of these telic higher-order thoughts therefore constitute the contents of the 'water' concepts of (P1) and (P2) identically.

Loar's Objection

Brian Loar objects to the Burge–Heil reconciliation story on the grounds that it cannot be used by *S* to construct a non-question-begging, wholly a priori justification of the pragmatic inference from the (presupposed) existence of Socrates and hemlock to *S*'s now thinking that Socrates drank some hemlock. To summarize briefly our earlier discussion: As a precondition of using the Burge–Heil reconciliation story, which assures *S* that all thoughts of a certain form are self-verifying, *S* would first have to know that he is thinking a thought of the relevant self-verifying form. In order to know, by way of applying the reconciliation story, that he is thinking "Socrates drank some hemlock," *S* would first have to know that he is thinking "I am thinking that Socrates drank some hemlock." But

given the coherence of the doubt that he stands in the appropriate external relations to Socrates and to hemlock, *S* would beg the question if he were simply to assume that he knows this a priori, without reliance on empirical observation of the world. I extended Loar's objection to include the charge that *S* would be launched on an infinite regress were he to attempt to construct a wholly a priori demonstration that he is thinking "I am thinking that Socrates drank some hemlock." And I suggested that the issues are substantially the same if we substitute for Loar's 'a priori' my favored terms 'direct[ly]' and 'authoritative[ly]'.

In responding to Loar, we should first distinguish two ways in which the Burge–Heil reconciliation story may be used. The reconciliation story may be used by theorists in constructing an account of *S*'s epistemic *entitlement* to his belief about his own thought. Alternatively, the reconciliation story may be used by *S* himself, as part of his epistemic *justification* of his belief about his own thought. In the latter case, *S* may actually run through a bit of explicit reasoning that includes the reconciliation story, or he may merely be disposed to use the reconciliation story in that way, and be epistemically justified in his self-knowledge in virtue of that disposition. But a correct account of *S*'s entitlement may be such that *S* is not even disposed to run it through; it may not be accessible to *S*, and if it is accessible, he may not agree with it.[17]

In what I shall call the base or *level-0* case, *S* does not use the Burge–Heil reconciliation story, nor need he be disposed to do so. Rather, that story is used by the theorist to explain *S*'s entitlement to his self-knowledge. *S* simply thinks a thought with content *p*. In so doing, *S* is a cognitive agent; his thinking that *p* is relevantly like an intentional act, and not merely something that happens to him. Thus *S* bears a telic relation to the content ⌜I think that *p*⌝. This spontaneously (in Hampshire's sense) yields a thetic relation to the content ⌜I think that *p*⌝, and this thetic relation is *S*'s authoritative self-knowledge. Because it arises spontaneously, it is noninferential. Yet *S* is entitled to it. The Burge–Heil reconciliation story, augmented by the above account of first-person authority as the product of the performative character of telic higher-order thoughts and Hampshire-style spontaneous belief, explains *S*'s entitlement.

If *S* realizes that he has self-knowledge, it may occur to him to wonder about its nature; he may wonder how his self-knowledge could be authoritative. This gives rise to the *level-1* case, in what will turn out to be a ladder of justifications. *S* may read Burge's and Heil's articles and may want to construct an explicit justification for his own self-knowledge. So, *S* proceeds: He thinks a thought with content, ⌜I think that *p*⌝.

So far we do not understand how *S* could have the basis for straightforwardly applying the Burge–Heil reconciliation story. For all *S* has is the thought, that he, *S*, thinks that *p*, and this higher-order thought, even if it is a belief, is not yet presented as something that *S* can think *about*. He can think *with* it, and use it as a premise in reasoning, but as a

premise in reasoning it does not interact with the Burge–Heil observation that thoughts of a certain form are self-verifying and are therefore true, since the thought that *S*'s thought is about only has the form *p*. From the thought ⌜I think that *p*⌝, *S* is not in a position straightforwardly to infer that he thinks any sort of reflexive or higher-order thought, and that is what he needs.

Yet intuitively it seems that *S* does have some sort of basis for applying the Burge–Heil reconciliation story, and his thought ⌜I think that *p*⌝ is key to the application. But that thought will not serve usefully here as a premise in reasoning. Rather, it is a performance, just as the thought that *p* was a performance in the level-0 case, and it is in some nonobservational sense on display, or perhaps better, it is the subject matter of reasoning, rather than a premise in reasoning.

The reader may perhaps have anticipated my account of the level-1 case: In thinking ⌜I think that *p*⌝, *S* is a cognitive agent. Thus *S* bears a telic relation to the content ⌜I think that I think that *p*⌝. This spontaneously (in Hampshire's sense) yields a thetic relation to the content ⌜I think that I think that *p*⌝. This is the relevant premise that figures in *S*'s application of the Burge–Heil reconciliation story. This premise constitutes authoritative noninferential self-knowledge, and we as theorists have explained *S*'s entitlement to it. From it, together with the Burge–Heil reconciliation story, *S* infers that his own reflexive thought is *true*. That is, *S* infers, ⌜My thought, I think that *p*, is true⌝; or, more simply, ⌜I think that *p*⌝. And this is the self-knowledge of first-order thought for which *S* has now supplied an explicit justification.

There is still, as always, an opening for a new philosophical anxiety. For *S* may realize, under the influence perhaps of reading Loar's article, especially the part quoted above, that his carefully constructed justification of his self-knowledge depends on his belief ⌜I think that I think that *p*⌝. Does this belief, perhaps, depend on some kind of nonauthoritative observational process? As theorists we have constructed an account of *S*'s entitlement, but of course *S* does not know that. So now *S* may wish to construct an explicit justification of this belief. Call this the *level-2* case. So, *S* proceeds: He thinks a thought with content, ⌜I think that I think that *p*⌝.

Again, *S* is in no position simply to use this current thought as a straightforward premise in explicit reasoning with the Burge–Heil account. If *S*'s goal is as stated, namely, to construct an explicit justification for the premise of the explicit reasoning of level-1, this current thought is of the wrong form to interact with the Burge–Heil account. If his (revised) goal is to take a short-cut, and reason from the current thought together with the Burge–Heil account to ⌜I think that *p*⌝, then he lacks a justification of his starting point, and we as theorists lack an account of his entitlement to it.

As before, we must think of the current thought as a performance, and not as a premise. Intentionally thinking the current thought, *S* bears a telic relation to ⌜I think that I think

that I think that p⌉, and spontaneously he bears a thetic relation to the same content. This is authoritative, noninferential self-knowledge, and we as theorists can explain his entitlement to it. S uses it as a premise, together with the Burge–Heil reconciliation account, to infer ⌈I think that I think that p⌉—the premise of the level-1 case. A second application of the Burge–Heil account allows S to infer ⌈I think that p⌉, which constitutes self-knowledge of first-order thought content.

If all that a reflective and epistemically cautious thinker S has at his command is the unadorned Burge–Heil observation that thoughts of a certain form are self-verifying, then he can never fully reassure himself of the directness and authority of his self-knowledge. For every time he climbs a rung of the ladder of explicit justifications, he lays himself open to a new Loar-inspired anxiety, focusing on the status of his new noninferential starting point, which will spur him to climb the ladder one more step, and so on ad infinitum.

As theorists we can construct an account of S's entitlement to his starting point at any level, including level-0 where S engages in no justificatory reasoning at all. So an unreflective thinker S does indeed have direct and authoritative self-knowledge, and so does a reflective thinker if his epistemic scruples are not so great as to cause him to renounce his birthright, so to speak, to direct and authoritative self-knowledge. But S can never fully reassure himself, by way of the unadorned Burge–Heil observation, that he has such direct and authoritative self-knowledge, for he will always anticipate a doubt attendant on the next rung of the ladder of justifications.

Nor will a reflective and cautious thinker be able to fully reassure himself if he simply adds to the unadorned Burge–Heil observation—unsupplemented by our current account—the distinction between entitlement and justification. If S is trying to reassure *himself*, then he is at once both thinker and theorist, and this blurs the distinction between entitlement and justification. What S requires is something in the *content* of the account of his entitlement, and not merely in the fact of its *being* an account of entitlement, as opposed to a justification, that will permit him, scrupulous as he is, to stop climbing the ladder of justifications.

The larger picture presented here, however, should give S the epistemic confidence he needs to stop climbing the ladder of justifications. He will realize that however high he climbs, he will inevitably have to start with a mental act. A mental act per se requires no epistemic justification, if its role in the argument is to supply the subject matter of thought, rather than a premise in reasoning. S will realize too that his thought about what his mental act thus supplies is authoritative because of both the performative character of his telic higher-order thought—what concept he exercises on an occasion is *up to him*—and the spontaneous belief he has of contents to which he stands in telic relation. So if S thinks of himself as a cognitive agent, and as entitled to a mental act as a starting point, he will see that further climbing of the ladder of justifications is entirely optional.

Acknowledgments

This paper first appeared in *Philosophical Topics*, vol. 24, no. 1 (spring 1996), pp. 71–99. Permission to reprint is hereby gratefully acknowledged.

For helpful comments on some of these ideas I thank Brad Armendt, Tyler Burge, Stewart Cohen, Brian Loar, and Steven Reynolds. Parts of this work were presented at a conference in October 1993 at Simon Fraser University on the work of Tyler Burge, at the University of California at Davis, and at the 1996 meeting of the American Philosophical Association, Central Division. I profited from the discussion on those occasions. This work was supported by a Faculty Grant-in-Aid from Arizona State University.

Notes

1. The representation that *p* may be thought of as a mental event of thinking that *p*, individuated by its intentional properties. I do not mean to be committed to a language-of-thought hypothesis, according to which representations are individuated by physical or syntactic features that are constituted prior to or independently of intentional properties.

2. See Putnam (1975), pp. 215–271, and Burge (1979), pp. 73–121.

3. See Burge (1988), pp. 649–663, and Heil (1988), pp. 238–251.

4. It should be noted that Loar puts knowledge of one's own references at the center of the topic of epistemic authority. He is primarily concerned with the question of how a thinker can make a priori inferences like:

Socrates exists ⊢ 'Socrates', as I use it, names Socrates.

This inference is, as Loar shows, pragmatically valid. That is, if an utterance of the premise is true in a certain context, then an utterance of the conclusion would be true in the same context, even though the propositions expressed by the premise and conclusion are related only contingently. Yet, Loar argues, a thinker who knows that this inference is pragmatically valid cannot use that knowledge to reassure himself that he can make the inference a priori. But rather than present Loar's argument on this point, I turn directly to a variant of the argument that Loar presents (on pp. 63–64) for our topic, the case of knowledge of one's own mental contents.

5. Loar is concerned with the apriority of a certain inference from existence premises; my focus is on the directness and authority of self-knowledge, which I see as typically noninferential. Moreover, as noted earlier, Loar's main concern is with knowledge of one's own references, which he sees as a more basic problem than knowledge of mental content; his argument against the Burge–Heil account is presented as a corollary consideration. Loar's own solution to the puzzles he raises requires that ordinary object-level concepts have reflexive implications. Space prohibits a critical discussion of Loar's views, but I take it to be an advantage of my solution to Loar's puzzle, presented below, that it does not treat ordinary object-level concepts as about themselves in any sense.

6. See also Boghossian (1994), pp. 33–50.

7. For the term 'tropistic' and related discussion, see Mark Johnston (1988), pp. 63–91.

8. See Shoemaker (1988), pp. 183–209; (1990), pp. 187–214; (1991), pp. 127–149.

9. Bas C. van Fraassen calls attention to Hampshire's view in van Fraassen (1995).

10. See P. F. Strawson (1959), p. 99, and Evans (1982), pp. 224–235.

11. For a defense of something like this view, see Loar (1988), pp. 99–110.

12. See my "Telic Higher-Order Thoughts and Moore's Paradox" (1995) for a critical discussion of David Rosenthal's HOT theory of consciousness. Note that for present purposes I require an occurrent unconscious

HOT only as a necessary condition, not as an explanation or analysis, of ordinary non-self-conscious conscious thought.

13. Twin Earth switching is not necessary to generate equivocal thoughts. Slow switching between America and England may suffice. Suppose *S* acquires the term 'endive' while growing up in America and subsequently moves to England, where 'endive' refers to a different vegetable (or so let us suppose; see Stich 1983, p. 63). Acculturation may occur over a period of decades without *S*'s learning anything about what the English called 'endive' that would distinguish it from what Americans call 'endive'. In that case, *S*'s thoughts containing his concept *endive* would be equivocal; they would have two intentional contents at once, without *S*'s knowing it—indeed, without *S* being in any position to know it a priori.

Moreover, in my view, there are many actual cases of equivocal thoughts that are generated by a thinker's simultaneous allegiance to both a scientific and a nonscientific linguistic community, cases where the thinker does not realize that the meanings of the relevant term or concept differ in the two communities. Think of someone who uses 'fruit' (do tomatoes count?) or 'nut' (do peanuts count?) and who maintains equal and simultaneous linguistic allegiance to the speech communities of both botanists and grocers. Such a thinker may think thoughts that express two propositions at once, without his being in any position to know this a priori.

14. In fact, I think, all hot self-knowledge is strong. Moreover, all smoldering self-knowledge at t_2 of what is in fact an ambiguous thought at (some earlier time) t_1 is also strong, unless further ambiguity in *S*'s thought has been introduced in the intervening period. Smoldering self-knowledge can be weakened only by the introduction of new conceptual ambiguity in the period between t_1 and t_2.

15. Boghossian seems to think that this is just what a violation of the principle of transparency of difference would have to consist in. This is a mistake; simpler violations are imaginable. Suppose *S* is fairly confident that 'chicory' and 'endive' are synonyms, that they refer to the same herb and express the same property. In fact they refer to different herbs, and on externalist views of mental content, *S*'s concept *chicory* has an intentional content distinct from that of his concept *endive*. No matter how rational and reflective *S* might be, he is in no position to know a priori that his concepts have distinct intentional contents. So much the worse, an externalist might say—ought to say, I think—for Boghossian's principle of transparency of difference. (I likewise reject Boghossian's principle of "transparency of sameness," which turns out to be equivalent to transparency of difference given some plausible assumptions. These matters deserve more discussion than I am able to give them here.) But the Pavarotti case mounts a serious challenge to externalism, involving as it does thought tokens of the same syntactic type.

16. Whether he *does* so say, in the relevant sense, is not a simple matter of asking him; see the discussion below about equivocation in argument.

17. For more on entitlement versus justification, see Burge (1993), pp. 457–488.

References

Anscombe, G. E. M. 1957. *Intention*. Oxford: Blackwell.

Boghossian, Paul Artin. 1989. Content and Self-Knowledge. *Philosophical Topics* 17: 5–26.

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

Burge, Tyler. 1979. Individualism and the Mental. *Midwest Studies in Philosophy*. 4: 73–121.

———. 1988. Individualism and Self-Knowledge. *Journal of Philosophy* 85(11): 649–663.

———. 1993. Content Preservation. *Philosophical Review* 102: 457–488.

Descartes, René. 1971. *The Principles of Philosophy*. In *Descartes: Philosophical Writings*, G. E. M. Anscombe and P. T. Geach (eds.). New York: Bobbs-Merrill.

Evans, Gareth. 1982. *The Varieties of Reference*. New York: Oxford University Press.

Hampshire, Stuart. 1975. *Freedom of the Individual*. Princeton, NJ: Princeton University Press.

Heil, John. 1988. Privileged Access. *Mind* 97: 238–251.

Humberstone, Lloyd. 1992. Direction of Fit. *Mind* 101: 59–83.

Hume, David. 1888. *A Treatise of Human Nature*. L. A. Selby-Bigge (ed.). Oxford: Oxford University Press.

Johnston, Mark. 1988. Self-Deception and the Nature of Mind. In *Perspectives on Self-Deception*, Brian P. McLaughlin and Amelie Oksenberg Rorty (eds.). Berkeley: University of California Press.

Kobes, Bernard W. 1995. Telic Higher-Order Thoughts and Moore's Paradox. *Philosophical Perspectives* 9: 291–312.

Lewis, David. 1982. Logic for Equivocators. *Noûs* 16: 431–441.

Loar, Brian. 1988. Social Content and Psychological Content. In *Contents of Thought*, Robert Grimm and Daniel Merrill (eds.). Tucson: University of Arizona Press.

———. 1994. Self-Interpretation and the Constitution of Reference. *Philosophical Perspectives* 8: 51–74.

Putnam, Hilary. 1975. The Meaning of 'Meaning'. *Mind, Language, and Reality: Philosophical Papers*, volume II. Cambridge: Cambridge University Press.

Shoemaker, Sydney. 1988. On Knowing One's Own Mind. *Philosophical Perspectives* 2: 183–209.

———. 1990. First-Person Access. *Philosophical Perspectives* 4: 187–214.

———. 1991. Rationality and Self-Consciousness. In *The Opened Curtain: A U.S.–Soviet Philosophy Summit*, Keith Lehrer and Ernest Sosa (eds.). Boulder: Westview Press.

Stich, Stephen. 1983. *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.

Strawson, P. F. 1959. *Individuals*. London: Methuen.

van Fraassen, Bas C. 1995. Belief and the Problem of Ulysses and the Sirens. *Philosophical Studies* 77: 7–37.

Woodfield, Andrew (ed.). 1982. *Thought and Content*. New York: Oxford University Press.