22 Mental Agency in Authoritative Self-Knowledge: Reply to Kobes

Bernard W. Kobes's creative and closely reasoned paper connects the performative element in basic self-knowledge to knowledge of what one will do in intentional action. I think this an illuminating comparison. In both cases, the knowledge derives from intellectual control over what one is doing, and in both cases the warrant derives partly from one's status as a rational agent. Some knowledge of what one will do is partly constitutive of being an intentional or rational agent. Some performative knowledge of one's propositional attitudes is partly constitutive of being a critically rational agent.[1]

I

There are differences between the cases. When one engages in intentional physical action, knowledge of what one will intentionally do depends partly, in a relatively immediate way, on matters outside one's control. One can know that one will pick up a fork, when one intends to do so. But the knowledge depends on the veridicality of one's perception of the fork and on one's using the perception to guide one's hands in the normal way to grasp the fork. Knowledge of one's thoughts in relevant performative cases is not hostage to the brute contingencies on which perception and physical action rely.

One can also know that one will raise one's arm. Here the perception of an object is not necessary in quite the way it is in the case of picking up a fork. One's perceptual relation to one's arm may be through proprioception. Proprioception is subject to fewer brute contingencies than visual perception. Still, I think that the dependence on a sense renders one liable to brute errors. In knowledge of what one will intentionally do through physical action, one is generally subject to such errors. One's intentions can be thwarted in ways that undermine true belief, hence knowledge, of what one will do. Some of these ways do not undermine one's epistemic warrant, nor do they involve malfunction of one's cognitive capacities. That is to say, brute error seems always to be possible. By contrast, authoritative self-knowledge of one's own mental states and events, including self-knowledge that involves a performative element, is not subject to brute error.

Self-knowledge that involves a performative element includes a broader range of cases than those that are *logically* self-verifying. The thought *I am hereby entertaining the thought that writing requires concentration* is logically self-verifying. I call such cases *pure cogito cases*. The logic and meaning of the intentional thought content requires that if the content is thought, it is true. But there are cases that normally involve self-verification and a performative element, even though the intentional content does not strictly entail that a thinking of the thought be self-verifying, or even true. These are *impure cogito cases*. For example, one can conceive of a case in which one commits oneself to the whole content of the thought *I am hereby thinking (in the sense of committing myself*

to the view) that writing requires concentration without committing oneself to the truth of the thought that writing requires concentration, thus without making the whole content (*I am thinking . . .*) true. One might believe that one is committing oneself to the component thought, but one simply absent-mindedly thinks it through. But such a case would be highly abnormal, even pathological. The intentional content is such that its normal use requires a performative, reflexive, self-verifying thought. It is an (impure) *cogito* case only where the thought is performatively self-verifying.

Or on writing out a check for a charity organization, one might think *I intend (*i.e., *hereby form the intention) to give to Oxfam*, where making the judgment about one's intention *is* the formation of the intention. The same act constitutes both. Thoughts of this form are fallible. Performative cases of this sort also count as impure *cogito* cases. One can not only think through this content and leave it false. One can even judge-true the content (*I intend* [i.e., *hereby form the intention*] *to give to Oxfam*) yet fail to form the intention, thus leaving the judgment false. But such a judgment would again be pathological. The normal use is to form the intention performatively in the formation of the judgment. So the act of making the judgment makes the judgment true.

In all of these pure and impure *cogito* cases, intellectual control is coordinate with the act of thinking a thought—whether forming a belief or making a judgment—that both commits one to the truth of the intentional content and carries out the thought that the content is about, thereby making the thought true.

I believe that most performative cases of authoritative self-knowledge involve a reflexive element. Although I think that there is something to Kobes's description of performative cases as 'telic', I am not sure that there is full agreement on these matters. Kobes's explanation of this notion is somewhat metaphorical. But it seems to me that in most paradigmatic cases of judgments with performative elements, including basic self-knowledge, the relation is reflexive or reciprocal between judgment and subject matter, rather than unidirectional. For example, when I judge, in the performative way, *I hereby form the intention to give to Oxfam*, as I make the decision and begin to write the check, the judgment is normally reflexive. It normally constitutes, rather than being caused by, the formation of the intention. In performative cases, a mental act of self-attribution makes itself true.[2]

Of course, the judgment about the intention *can* be caused by an antecedent intention. The intention can form just beforehand; or it can be a standing state. (Or perhaps the causation can be simultaneous, but between distinct events.) I believe that self-attributions of such intentions and beliefs can be authoritative. But in most cases they are not, in my view, strictly performative. Since cases of authoritative self-knowledge in which the attributed propositional attitudes are occurrent thoughts or intentions that cause the self-attribution are similar to cases of authoritative self-knowledge of standing propositional attitudes, I

shall discuss the two types later under the rubric of authoritative self-knowledge of standing states.

Kobes holds (p. 204) that we have self-knowledge of what we are thinking even when we are thinking ordinary conscious thoughts about the world, and that in many such cases we are not thinking "explicitly" reflexive thoughts. I agree that self-conscious self-knowledge is present in many ordinary first-level thoughts about the world. But I think that there is a reflexive element in more such self-knowledge than most people realize. I do not know what he means by 'explicitly'. We do not often verbalize such thoughts with 'hereby' or 'in this very thought'. But I think that when we have authoritative propositional self-knowledge of what we are presently thinking when we are thinking about the world, particularly when there is a performative element in the self-knowledge, there is normally—or at any rate, very frequently—a reflexive second-order element in the logical form of the first-order thoughts.[3] Although these performative or reflexive cases form a larger class than one might first think, I believe that they do not constitute the whole of what I see as authoritative self-knowledge. Kobes places heavy emphasis on the activity of belief-formation. I agree that agency is at the heart of our understanding of first-person authority. But I doubt that it can bear weight in just the way that Kobes or Descartes require it to.

Here Spinoza seems to me to provide a salutary qualification on the Cartesian view to which Kobes's emphasis is congenial. Descartes maintained that belief is always the product of an act—in fact a "willed" act—to assent to a proposition. Descartes seems to me right in maintaining that some instances of belief-formation are instances of a type of agency, and some of these cases are subject to considerations of intellectual responsibility. But I think that this model should not be fully generalized. Opposing Descartes, Spinoza maintained that belief is the default position, not an activity on a proposition that one noncommittally understands. He maintained, as a thesis of philosophical psychology, that belief is concomitant with understanding. Spinoza held that doubt and suspension of belief are acts. One can undo the initial nonactive default position. But formation of belief is, on his view, not an act. It is automatic if not checked.[4]

Spinoza's view, in its fully general even if vague form, seems exaggerated. But he seems to me right in maintaining that the formation of a great number of beliefs, particularly perceptual beliefs, is not strictly an activity by the whole mental agent. Some of our authoritative self-knowledge resides in our self-attribution of such beliefs. These considerations help motivate my view that the performative model cannot fully explain first-person authority.[5]

Let us lay aside modular and otherwise inaccessible beliefs. There are still beliefs and other propositional attitudes whose formations do not constitute acts of ours. They seem rather to form in us. They may be part of a functional organization that is essential to being

an agent. But not all such beliefs that are accessible to self-conscious self-attribution are themselves the products of acts of commitment. At least, I can think of no natural sense of agency that applies to their formation. Most perceptual beliefs are of this sort. Many beliefs that rest on interlocution, especially in childhood, are as well. Beliefs that derive from perceptual beliefs by way of hard-wired inductive mechanisms—as opposed to active, person-level inferences—are also not in any obvious sense products of agency.

Kobes sometimes speaks of nonmodular beliefs as "up to us." There may be something in this. They make up a point of view that we as doxastic agents can claim as ours. Once we become critical reasoners, we are epistemically responsible for them. But it does not follow that they are products of agency. Rather, we may be responsible for revising them if counter-considerations arise, and we may have a parallel responsibility for their maintenance. It would seem to me mistaken to hold that all nonmodular propositional attitudes—or even all attitudes that we are potentially authoritative about—are formed through intellectual agency. Thus I think that our authority about occurrent and standing attitudes is not captured fully by appealing to the model of "hot" or "smoldering" telic self-knowledge.

Nor do all cases of authoritative self-knowledge exhibit an ability to bestow the content on the propositional attitudes that form the topic of the knowledge. Often those attitudes have a nature and existence that is prior to and independent of authoritative judgments about them. Authoritative knowledge of our standing or occurrent perceptual beliefs, and of many of our other standing propositional attitudes, needs, I think, a broader account.

Of course, in authoritatively attributing propositional attitudes to ourselves, we normally commit ourselves to those attitudes. Claiming them as one's own may seem in effect to endorse them, at least when they are not viewed as foreign "objects" inside one's self. And in those "foreign object" cases, the self-attribution is not authoritative. This may suggest that agency is pervasive in the propositional attitudes thus self-attributed. I think that this is an important point.

But the truth of some authoritative self-attributions does not depend on these endorsements. In making some authoritative self-attributions, one is making a judgment that is not intended to be made true by one's endorsing the attitude on the spot. It is intended to capture a stable attitude that was present antecedent to the self-attribution, and whatever re-endorsement that might involve. As example, consider: *I believe that my sister is younger than I am*. Such self-attributions are fallible. The ways that they are corrected—for example, by reference to past statements or behavior—show that endorsements of the first-level propositional attitude that are implicit in the self-attributions themselves are not in general taken to be sufficient to guarantee the truth of the self-attributions.

Thus a full account of the specialness and authority of some of our self-knowledge needs to go beyond both Kobes's telic model and beyond my paradigm of self-verification and

performative acts in *cogito* cases, pure or impure. When I introduced my self-verifying model as a paradigm, I called the sort of self-knowledge that it encompasses "basic self-knowledge." But I was fully aware, as Kobes recognizes, that there is a range of other cases of authoritative self-knowledge that does not exhibit self-verification. There is first-person knowledge of sensations, of occurrent perceptual beliefs, of nearly all standing states that predate the formation of judgments or even knowledge about them. There are certain cases of memory. There is knowledge of some of one's feelings or emotions. And so on.[6]

II

My view in "Individualism and Self-Knowledge" was that there are features of self-knowledge—other than self-verification—that are dramatically and paradigmatically realized in what I called the basic cases and that provide a key to understanding the whole range. I have yet to carry out this strategy fully. But I have taken some further steps.[7] And I still believe that it is a viable and promising enterprise.

I want to stress that "Individualism and Self-Knowledge" was not intended as a full account of authoritative self-knowledge. Not only does it indicate that the paradigmatic "basic" cases do not constitute the full range of authoritative self-knowledge. It is focused on the semantical or logical role of attribution of intentional content in self-knowledge and on the easiest cases of a self-verifying attitude relation. It does not discuss the notion of warrant, in any depth, *even for the basic cases*.

The point of the article was to raise, and provide an initial response to, an apparent problem about the relation between authoritative self-knowledge and anti-individualism as an account of the nature of propositional attitudes. Although I made a number of comments about differences between authoritative self-knowledge and perceptual knowledge, I concentrated on pointing in a direction for an account of authoritative self-knowledge. Criticism of the paper for being "thin" as an account of authoritative self-knowledge is off the mark. The points, which Kobes cites others as making, about differences between the basic cases and other cases of authoritative self-knowledge were points that I anticipated and in several cases explicitly noted myself. I thought that those were matters to be taken up later. I still expect to deliver on the promissory note. I hope that this reply will advance matters.

Kobes's discussion of extensions from the paradigmatic cases is sympathetic and discerning. I think that he is right that some cases of authoritative memory and some cases of authoritative knowledge of standing states can be understood in terms of mechanisms preserving traces of earlier states known or knowable in the paradigmatic ways.[8] I think that these mechanisms should be seen as part of the rational apparatus that is constitutive

of being a critical reasoner. As I have noted, however, I do not think that all of these cases of preservation can be assimilated to the act-paradigm that Kobes develops.

III

I turn now to the two objections Kobes discusses. His treatments of the Loar and slow-switching cases are well reasoned, and broadly plausible to me. My views differ in some significant ways, however. There are also some matters I would like to try to clarify.

Kobes's discussion of slow-switching cases makes some points in common with my discussion of such cases in my (1998b).[9] In that recent paper I focused on unaware slow switching rather than knowledgeable slow switching. I emphasized the nonreflective role of preservative memory in knowledge of one's past thoughts.[10] Even in knowledgeable slow switching, I see no reason to think that the individual cannot remember the past thoughts, even if he cannot distinguish them from cohabiting twin thoughts. The individual is not forced to ask the question that Kobes and Boghossian focus on. He need not ask whether he was thinking about water or twater. In fact, where he asks this question in this way, I believe that he loses his authority in his application of memory (see Burge 1998b).

The individual can instead rely on preservative memory to take up the content and attitude-type of the thought that he in fact thought at the earlier time. Relying on memory to individuate rather than to preserve would be a mistake. Such reliance in the switching cases could indeed undermine knowledge of the past. It would treat the remembered event as an object, rather than anaphorically—as part of a single point of view. In such a case, the memory would not be authoritative.

If one uses information about having been switched between earth and Twin Earth to try, through memory, to discriminate earth- from Twin Earth-type thoughts, one loses one's authority over one's past thoughts. Errors deriving from such uses of memory are not naturally counted cases of forgetting. But if they are not so counted and are seen rather as unfortunate use of newly acquired information, then, as Kobes and I both point out, Boghossian's "platitude" that if one forgets nothing one cannot lose knowledge is shown to be false. Indeed, the "platitude" is false for a wide range of cases, many of which have nothing to do with switching scenarios. New information, misleading information, can drive out old knowledge. So no pressure is generated on the anti-individualist by these means.

I am uneasy about Kobes's appeal to equivocal thoughts as the entire basis for his account of reasoning in the relevant switching cases. He is, I think, quite right to reject the idea that anti-individualism commits us to an unacceptable susceptibility to equivocation in deductive reasoning, even by the most rational reasoners. It may be that in some instances one thinks equivocal thoughts. Or as I would prefer to see it, one may think mul-

tiple thoughts, on given occasions, without distinguishing them. It may be that Kobes's idea is part of a necessary solution. But all of the reasoning cases that I know of can be accounted for by noting that even someone who has switched and has "twin" concepts in his repertoire is usually not using both concepts when he uses either of them. Rather, features about the cognitive context and cognitive point of a thought determine which of the concepts is employed. I refer the interested reader to a fuller account of the matter in my (1998b).[11]

Thus I do not accept Kobes's assumption (p. 217) that because the switched individual is on Twin Earth and has the twin concept, he "cannot escape including an exercise of the Twin Earth concept" in any given thought in which he remembers something about the twin object back on Earth. This assumption suggests some magical effect that merely being on Twin Earth has on one's thinking Twin Earth thoughts (once one has acquired them). I believe that, at least in the cases that have been discussed in the literature so far, one does not have to appeal to equivocal or multiple thoughts to block unacceptable results about reasoning in slow-switching cases.

An analogue of the reasoning cases occurs for demonstratives, without appeal to anything as complex as Twin Earth. A person can be looking at a tomato and think *that is healthy so that is healthy*. In place of a healthy tomato, a perceptually indiscernible, rotten one could be substituted so quickly between the antecedent and the consequent of the thought that the reasoner would not notice. Then a rational agent *could* become committed to an invalid proposition through no irrationality—if he used the demonstrative twice and independently to indicate the object before him. He might even assume mistakenly that the thought is a logical truth.

A correct account of the logical form of the invalid, false thought will not treat 'that' in its two occurrences as being syntactically the same. The token applications count, from the point of view of a syntax relevant to logical form, as formally different. If the thinker treats both occurrences as deictic, as independently applied demonstratives, then he is liable to error with regard to what could *seem* to be a very safe logical truth. But the thinker would not be committed to the syntactic or logical form of a logical truth. If the thinker does not treat both occurrences as deictic, then there need be no difficulty. If the thinker is to avoid any susceptibility to difficulty, the second occurrence of 'that' must be tied anaphorically to the first. Obviously, one could avoid switching the tomatoes, and still get the contrast between the logical truth and the non–logical truth. The difference does not lie in the objects or *res*. The difference lies in the uses or applications in thought that are counterparts of linguistic uses of the demonstrative 'that'. These are under the potential cognitive control of the thinker.

Although the twin concepts that I have discussed in slow-switching cases are not demonstrative or indexical, the twin concepts contribute differently to logical form or logical

syntax. One can engage in equivocation without realizing it, in a way analogous to the way one can engage in a demonstrative shift of reference without realizing it.

To avoid susceptibility to equivocation, one must tie one's concepts together "anaphorically" in one's reasoning. But this is the normal way that we implicitly understand steps in a piece of deductive reasoning as fitting together anyway. We implicitly understand such steps in that way even apart from anti-individualist considerations. When I think "Every man is mortal; Socrates is a man; so Socrates is mortal," I allow no equivocation on 'man', 'mortal', or 'Socrates' by implicitly relying on a sameness of conceptual and indexical use.

If I understand him correctly, Kobes makes substantially this point (pp. 211–212). But I think that the links between steps in reasoning are formed anaphorically. Sometimes there is a telic line that is "forged," but it seems to me that commonly the anaphora is correctly seen as simply preserving a content that was already unequivocally in place at the earlier step in the reasoning. Such preservation seems more "thetic" than "telic," but I am quite happy to dispense with this terminology.

I turn now to Kobes's discussion of Loar's doubt. There are elements in the very posing of the doubt that seem to me to be odd and in some respects significantly off the mark. The problem is supposed to be to explain how someone could assure 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*.

In the first place, the relevance of the existence assumptions seems to me quite unclear. Kobes (on Loar's behalf) writes, "Now let us suppose that *S* knows that such reflexive thoughts [as "I am now thinking that Socrates drank some hemlock"] are always true, given the existence presuppositions [that Socrates and hemlock exist]" (p. 203). The authority of authoritative self-knowledge does not extend to the *res* in *de re* judgments. The relevant self-attributions in authoritative self-knowledge are to be seen as about the intentional content and the attitude-type of the attributed attitude. There are intentional elements in thought referring to Socrates and hemlock. These are trivially not identical with Socrates or hemlock. They have aboutness or representational properties and functions. Socrates and hemlock do not. The intentional or representational content does not even *include* the referents of the conceptual elements. It includes only the "senses" or modes of presentation, or conceptual and applicational elements of the thoughts.

Thus it is epistemically possible that one think mistakenly that Socrates drank some hemlock even if Socrates and hemlock were to turn out not to exist. If we were to find out that Socrates did not exist, I would still have thought a thought properly expressed in such terms. And I would have authoritative self-knowledge of my thought (of the thought that I would have thought) even if Socrates and hemlock were to turn out not to exist. In that case, my thinking the relevant thought would not, of course, depend on my bearing causal

relations to Socrates himself or to hemlock. It would depend on another causal complex connecting me to other relevant uses of the name 'Socrates' or the noun 'hemlock'.[12] Given that Socrates and hemlock do exist, of course, my thinking the intentional content of the thought—not merely my thinking things *de re* about Socrates—depends on my actually bearing those causal relations to those objects. But from an epistemic as opposed to individuative point of view—which is the point of view at issue—the existence or nonexistence of Socrates and hemlock is irrelevant to the self-knowledge.

In the second place, and more centrally, it seems to me that the challenge to produce an apriori *inference* is misconceived. Kobes takes up this challenge to show how a thinker *S* "could demonstrate to himself a priori that he is thinking that Socrates drank some hemlock." But I think that the challenge itself needs some scrutiny.

Loar's idea appears to be that, in light of externalist claims that thinking the reflexive thought depends on bearing causal relations to some external objects, one should worry that one could know one's thoughts only empirically, because one can know these relations to the environment only empirically. So one needs to be able to reproduce some apriori inference from the existence presuppositions to the reflexive thought.

That this way of understanding the problem that I set up in "Individualism and Self-Knowledge" is offtrack is suggested by the following fact. No inference to a conclusion constituted by one of the reflexive, self-verifying *cogito*-cases could possibly be needed to justify the relevant thought. Consider the thought *I am in this very thought entertaining the thought that Socrates drank some hemlock*. Because of the performative, self-verifying character of the thought, no inference to it could possibly be needed to provide support for it. It is clearly a starting point, a basic thought, whose justification lies in its own performance and content, not in the content of other thoughts from which it might be inferred. So no inference to it is needed for it to have epistemic support. This point seems to me to be completely independent of the truth or falsity of anti-individualism. So the demand that one be able to support it through some inference is misconceived.

Of course, not all instances of authoritative self-knowledge are reflexive, performative, or self-verifying. In fact, the real issue is not one of justifying the truth of the whole self-attribution. It concerns the warrant for the attribution of the *intentional content* of the attitude that one attributes.

This remark brings me to a third respect in which the problem is misconceived as posed. In my (1988), I asked why the fact that we have only empirical access to causal relations that fix the nature of our thoughts does not entail that we cannot know that we are thinking such and such unless we engage in empirical investigation that shows the conditions for thinking such and such are satisfied. I said that the answer "can be seen as a series of variations on the point that one must start somewhere" (Burge 1988, p. 654).

I think that Loar and Kobes overlook the starting point that I took to be basic. They suggest that the key to my "reconciliation strategy" is a schematic generalization: Kobes writes:

> What *S* knows is that all thoughts of a certain form—including "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. (p. 204)

He later calls the schematic generalization my reconciliation strategy "unadorned."

There are two things wrong with this account of my view. One is that the *fundamental* relevant generalization is *not* that thoughts of a certain form are true. Only pure *cogito* cases—not even all performative or reflexive cases—of authoritative self-knowledge are true in virtue of their form, together with the fact that they are actually thought. For example, the thought *I hereby intend to give to Oxfam* is not true in virtue of its form, as *I am hereby entertaining the thought that writing requires concentration* is. Only the latter is a *pure cogito* case. I discussed both sorts of cases in my article. But the key generalization fixes on the intentional content attributed in the relevant thoughts, not their truth. Thus, the main issue concerns the contents *to give to Oxfam* and *that writing requires concentration*. The point, as I stated it, is that performative or reflexive cases are such that the intentional content that they attribute is thought and thought about at the same time. So the content of the attributed bottom-level attitude and the content attributed in the self-attributional thought are locked together. Thoughts of that reflexive form cannot mistake the intentional content of the attributed thought, though some of them—the performative cases that are not pure *cogito* cases—are fallible. Thus one could be mistaken in holding that one is *intending* in the relevant case. But one could not be mistaken because of some mismatch in content between self-attribution and an attributed intention. I think that one can know, by simply understanding one's thought, when one's thoughts have a reflexive form and require the relevant locking.

The apparent threat in the switching cases is that the content of one's self-attribution and the content of the attributed propositional attitude will come apart. The schematic generalization shows that this apparent possibility is illusory, at least in a large number of cases of authoritative knowledge. It is illusory in all *cogito* cases, whether pure or impure—all cases of reflexive performatives. Successful reflexive performatives, whether pure *cogito* cases or other sorts of reflexive performatives, are self-verifying. Some thoughts with the form of reflexive performatives—impure *cogito* cases—can be false. Yet in all such cases, the content attributed in the self-attribution cannot fail to be the content of the attitude that is attributed. One may mistake the attitude. But one cannot get the atti-

tude right and mistake its content—at least not in the performative, reflexive cases. More will need to be said about authoritative instances of self-knowledge that are not performative or reflexive, as for example, self-attributions of standing states. I shall return to these cases in a short while.

The second thing wrong with the account of my reconciliation strategy is more fundamental. The exposition Kobes gives of my account understates what my "reconciliation story" says. My account does not merely appeal to a schematic generalization. It emphasizes that people must understand the content well enough to think and attribute it. In particular cases of thinking the relevant performative or reflexive thoughts, the content is thought and thought about at the same time. The same content is deployed at the lower level and at the higher level of self-attribution at the same time.

I wrote that my answer to the problem I raised would be a series of variations on the theme that one must start somewhere. The basic starting point that I alluded to is one's understanding of the intentional content of one's own thought. The starting point is not a generalization about the form of thoughts. It is the minimal understanding necessary to think the self-attribution, and to raise sceptical scenarios with respect to it, in the first place.[13]

Thus thinkers who self-attribute in the relevant way must be taken to understand their contents well enough to think them. In thinking them in the reflexive, performative cases that I centered on, they must think and think about their contents in the self-attribution itself in such a way that the intentional content at the different levels cannot come apart. It is the same, understood content at both levels.

The problem as Loar poses it ignores the reflexive nature of the relevant thinking. By treating the thinker as if he must wonder what content he is thinking (in the conditional, "If I am now thinking that I am now thinking that Socrates drank some hemlock . . . ," p. 204), Loar treats the thinker's understanding of his own thought as if it were understanding from the third-person point of view. The thinker is not in a position to wonder about the content of his thought, in the relevant way, given that he makes the self-attribution. He must minimally understand the content in order to think the self-attribution in the first place. The problem as posed fails to acknowledge that the thinker must be able to understand, grasp, the particular intentional content in thinking it. In thinking reflexively, the thinker thinks the content and self-attributes it in the way that is expressed linguistically in that-clauses.

Kobes is right (p. 222) to make the important point that my account of reconciliation is from the point of view of a theorist constructing an account of the thinker's epistemic *entitlement* to the self-attribution. Or rather I am constructing an account of one element in the entitlement. An entitlement is an epistemic warrant that the individual has but need not have the concepts or abilities to explain or understand, even on reflection. So in my

view there is no need for the individual to be able to give an apriori, or any other kind of, account of why he is nonempirically warranted in his self-attributions. The individual need know nothing of anti-individualism or of any reconciliation strategy to have the relevant warrant.

In thinking that *I am now (in this very act of thought) thinking that Socrates drank hemlock*, I understand the content of my thought well enough to satisfy the condition on understanding that is relevant to making it a reasonable question whether the thought that I am thinking constitutes knowledge. Given that the understanding is in place, and given that I actually think the relevant thought, I cannot be mistaken in my self-attribution of the content. No empirical knowledge is needed to establish an understanding of the intentional content of the thought that one is thinking.

But suppose, as Loar and Kobes do, that the individual knows about anti-individualism and the reconciliation strategy. What are we to say about the challenge to produce an apriori inference to the self-attribution as conclusion? What I think we should say depends on the particular type of authoritative self-knowledge.

In pure *cogito* cases, no justificatory inference is possible, and none is needed. The judgment *I am hereby entertaining a thought that Socrates drank some hemlock*, when made, is self-evidently self-verifying. It is obvious that empirical issues and issues about switching are irrelevant. For from the point of view of the person making the judgment, the thought is understood by him and understood to be self-verifying, hence obviously true. It is at least as much a piece of nonempirical knowledge as is a self-evident truth of logic.

In cases of reflexive performatives that are not pure *cogito* cases—like *I hereby judge that Socrates drank some hemlock* or *I hereby intend to give to Oxfam*—the self-verification is not formally guaranteed merely by thinking the thought. But their truth is guaranteed, in normal circumstances, by the clearly understood performance of the act. As I indicated earlier, however, the mistake that is supposed to be threatened by the slow switching cases is not just any sort of mistake. It is supposed to be a mistake that derives specifically from some dislocation of content between that of the self-attribution and that of the attributed attitude. But again, no justificatory inference here is necessary, or, as far as I can see, possible. If one understands the content of one's mental act of self-attribution, one understands that there is no room for a transition between the content of the self-attribution and that of the attributed attitude. For the content of the attributed attitude is fixed as that of the self-attribution. One can think the first thought and fail to make a judgment, and thus think something false. One can make the judgment articulated in the second thought and fail to have the relevant intention, and thus judge falsely. But one cannot indicate through the thought a judgment or intention that lacks the content that is attributed—the intentional content: *that Socrates drank some hemlock*. One can understand these points on reflection. That is what the reconciliation point emphasizes. It simply

elicits something present in one's ordinary understanding. Again, empirical considerations and issues about switching are obviously irrelevant, once one reflects on one's own thoughts. So again no apriori inference is needed, or as far as I can see, possible.

It seems to me that Kobes may himself fail to appreciate the implications of reflexivity when he writes:

> 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. . . . 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. (pp. 222–223)

This seems wrong, or at least misleading. At any rate, it is wrong for the cases that I explicitly discussed, which are all reflexive, dual-level ('hereby', 'in this very thought') cases—performative analogues of the *cogito*. The relevant thought is *I think with this very thought that p*. Insofar as the performance is reflexive, as the cases I discussed explicitly were, at least the content is both thought and thought about in the same act. In the pure *cogito* cases (*I am hereby entertaining the thought that . . .*), the attitude relation *and* the content of the attitude attributed are both in the position of being both thought and thought about. This is true even in the nonpure reflexive case *I hereby judge that. . . .* In many of the other reflexive cases—for example, the intending case—at least one of the attitude relations lacks this dual-level role. Thus, the judgment about the intention is not being thought about as a judgment. But in all such cases, the intentional content that is attributed to a propositional attitude has the dual-level role. Only the intentional content is really at issue in the switching scenarios.

As far as I can see, for reflexive cases no hierarchy is relevant to the justification. One needs no engagement with a hierarchy to justify to oneself that one has not made a mistake that results from a disengagement of content between the self-attribution and the attributed attitude. And no inference is appropriate. It is self-evident from the understanding present in one's making the judgment itself that the problem cannot arise.

In these cases, the very posing of the problem results from ignoring the implications of the sort of reflexive understanding necessary to thinking the relevant intentional contents. The puzzle misleads one into treating those contents as objects of identification or potential investigation, or as otherwise separable from the content of the self-attribution. But in fact they are already understood and "individualized" in the only way necessary for the relevant self-knowledge.

The thoughts at both the self-attribution level and the attributed level are dependent for their content on environmental relations. But given that one is thinking the thoughts—and understands them *in* thinking them, as opposed to understanding them through empirical investigation or in some explicatory way—no switching and nothing about the

environmental relations could possibly lead one into error. This realization is not a product of an inference from premises. It is the result of reflection on the nature of one's understanding.

I have left open whether some self-knowledge that can reasonably be counted "performative" might not be reflexive. Kobes's solution goes through an ingenious discussion of how a hierarchy is generated through the telic mechanism. Perhaps sometimes a hierarchy *is* generated in some nonreflexive performative cases. I would like to understand this better. If there are such cases, I think his solution is likely to be, in its main outline, correct. I do not reject Kobes's account as wholly inapplicable to the problem. There *might* be cases that can be handled in Kobes's way at any level up an infinite hierarchy. But I do not think that the account covers the most common cases, or the cases that I centered my discussion on. In cases where one thinks the intentional content reflexively as the content of the attribution and of the attributed state or act, no hierarchy arises in dealing with the switching cases.

The explicit articulation of 'hereby' or 'in this very thought' is not necessary for reflexivity. Most self-attributions of occurrent attitudes are reflexive. Moreover, many cases of self-aware conscious thoughts that *p*—that is, self-aware occurrent propositional attitudes that do not explicitly formulate a self-attribution of the propositional attitude—are nevertheless reflexive self-attributions of the relevant attitude toward the content that *p*. The self-awareness often involves an unarticulated reflexive self-attribution. In judging consciously and explicitly that *p*, at whatever level, one commonly implicitly believes that one thereby judges that *p*, as a component of the judgment that *p*. Whether one also believes that one believes that one judges that *p* is a matter of the subtlety of one's self-awareness; it is not required by the self-awareness of the bottom-level judgment. Language often suggests a hierarchical separation that is not present in the actual thinking.

IV

Everything that I have said so far about the problem Kobes poses centers on reflexive cases, or at least performative cases. But I have emphasized that some authoritative self-knowledge is not performative or reflexive at all. I have in mind knowledge of one's standing states and of certain of one's past standing states—for example, perceptual beliefs—through preservative memory. In these cases, there is nothing in merely understanding the self-attribution itself that prevents a disengagement between the content of the self-attribution and the content of the attributed state. Is an apriori inference needed in these cases?

No, an entitlement to self-knowledge holds in these cases as well. So no inferential justification by the subject is *needed*.

But suppose the subject were apprised of anti-individualism and of the reconciliation strategy. Is a nonempirical justificatory inference possible?

It seems to me that an inference is not in place even in these cases. One needs to explain to oneself a relation of noninferential transition between lower-level attitude and self-attribution that is not subject to environmental vicissitudes or in need of empirical support. One starts by postulating some belief with a certain content—say, that Roberto drank some hemlock. This belief may not be the product of any mental *act*. It might be acquired perceptually, or through interlocution, in a nonactive way. It may or may not be warranted. Suppose that this belief has been residing in one. Something brings this belief to consciousness; or something causes one to remember the belief. One comes to employ the first-person concept and one's concept of belief in judging: *I believe (or remember) that Roberto drank some hemlock*.

Why, in light of the way that one's concepts of Roberto, drinking, and hemlock depend on relations to an external environment, is one entitled nonempirically to one's self-attribution? There can, in these cases, be dislocations between the content of one's initial standing state and the content of one's self-attribution. But insofar as the relation between standing state and self-attribution is not dependent on investigation of or other reliance on the environment, beyond the causal input that made the standing state with its particular content possible in the first place, no dislocation would be affected by switches. The self-attribution would simply inherit the content of the environmentally determined standing state.

I have elsewhere explained rational noninferential relations that fit this description. In the case of self-attribution of a present standing belief, the relation is necessary for rational deliberation. Individuals come to be reliable in making self-attributions through such relations. In the case of authoritative self-attributions of past standing states, the relation is a combination of purely preservative memory and the bringing to consciousness of the standing state that I just described. Purely preservative memory is the more basic rational relation of the two. It is necessary for engaging in any kind of reasoning in time, not just self-conscious critical reasoning. It preserves the content between different attitudinal states over time.

Both relations are or are supported by causal relations. Both can be broken in ways that would lead to error. But the breaks would be internal to the thinker's cognitive system. They would not involve brute errors. They would not in any way depend, for their accuracy in preserving the content of the attributed stated, on relations to or input from the environment. They are rational relations internal to a cognitive point of view and practice. Only the initial standing state, at least in the case of the belief involving Roberto, would depend for its nature, content, veridicality on relations to the environment. That content would then be inherited and operated on in the rational relations on which self-attributions

are founded. Similar points could be made in moving from first-level, standing self-attributions to mental acts that self-attribute, at the second level, the standing first-level self-attributions. And so on.

I have elsewhere characterized both such relations as essentially transitions *within* a point of view.[14] In this respect, they are like inferences. The transitions do not require that the initial subject matter for the self-attribution be mental acts. No reference is made to performatives in explaining them. There is no element of self-verification. The self-attributions are acts, to be sure. But they conform to an antecedently established subject matter. They are not established reflexively in the self-attribution itself. They are fallible. They may be thrown offtrack by bias, top-down reasoning, internal malfunction. But in cases of *authoritative* self-attributions they are not subject to brute error. Their relation to the content of the standing states is not hostage to vicissitudes of the environment. Their warrants derive not from relations to the environment that could be known only empirically. Their warrants derive from the reliability of the causal connections and from the roles of the relevant relations, and the associated self-attributions, in various aspects of rational systems. The warrants are thus not perceptually based. They are nonempirical.

Notes

1. See my (1996), pp. 91–116. The notion of brute error that is used in what follows was introduced in my (1988), pp. 649–663. Some of the points made regarding the relation between authoritative self-knowledge and brute error are also made in the same article.

2. I think that such performatives as *I promise to give it to you* are also self-verifying. I reject analyses that claim that they lack a truth-value.

3. See elaboration of this point in my discussion of examples (1) and (2) in my (1996).

4. Descartes, *Meditation* IV; Spinoza, *Ethics* II, 49. For psychological evidence bearing on the matter, see Gilbert (1991), pp. 107–119; and Wegner and Pennebaker (1993).

5. I believe that Kobes's brief remarks about our being the author of our thoughts (p. 208) tend to overrate our authorship. Many of our ordinary beliefs are not self-conscious or the products of agency, but we can be authoritative in our reflective self-attributions of them.

6. Although all of these extended cases involve the possibility of fallible self-attributions, I do not believe that they are any less authoritative than the performative or self-verifying cases, in my sense of 'authoritative'. There is at least the appearance of disagreement with Kobes on this point. See p. 215. I might also say that I find the talk of substantiality and insubstantiality, which derives from a very odd and misleading technical use of this term in Boghossian's article, unfruitful in many ways. See Boghossian (1989), pp. 5–26, section III. I regard authoritative self-knowledge, even in the self-verifying cases, as *substantial* in all normal senses of the word. Kobes's explanations of the term are different from Boghossian's, and I have no criticism of what he says on this score—except insofar as use of the term, with yet more different special meanings, remains a source of possible confusion.

7. Burge (1988). For the further steps, see my (1996), (1998b), (1998c), and (2000).

8. See my (1993), pp. 457–488; (1998a), pp. 1–37; (1998b); (1999).

9. Neither of us had read the other's account.

10. Kobes's "smoldering self-knowledge" is I think a special case of preservative memory—the case where the antecedent involves a performative. I also see the role of agency in preservative memory somewhat differently. But I believe that we are on to the same phenomenon.

11. The fundamental point that I make in this article is anticipated by Schiffer (1992), in his comment on the paper by Boghossian that Kobes discusses. Of course, reference to the contextual determination of which thought is thought is only part of the full account.

12. Such cases are discussed in my (1983), pp. 79–110.

13. In the last paragraph of his essay, Kobes seems to me to state very well an essential aspect of the starting point—that one starts with a mental act. My account emphasizes this as well, but also emphasizes the type of understanding commonly involved in performative cases, the reflexive understanding that is present in most performative-type self-attributions. The intentional content associated with the that-clause is thought and thought about at the same time.

14. See my (1996) and (1999).

References

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

Burge, Tyler. 1983. Russell's Problem and Intentional Identity. In *Agent, Language, and the Structure of the World*, Tomberlin (ed.), pp. 79–110. Indianapolis: Hackett.

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

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

———. 1996. Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society* 96: 91–116.

———. 1998a. Computer Proof, Apriori Knowledge, and Other Minds. *Philosophical Perspectives* 12: 1–37.

———. 1998b. Memory and Self-Knowledge. In *Externalism and Self-Knowledge*, Ludlow and Martin (eds.). Stanford: CSLI Publications.

———. 1998c. Reason and the First-Person. In *Knowing Our Own Minds*, Crispin Wright, Barry C. Smith, and Cynthia Macdonald (eds.). Oxford: Clarendon Press.

———. 1999. A Century of Deflation and a Moment about Self-Knowledge. Presidential Address at Pacific APA, April 1999. *Proceedings of the APA* 73: 25–46.

Descartes, Rene. 1641. Meditations in *The Philosophical Writings of Descartes*. Cottingham, Stoothoff, and Murdoch (eds.). Cambridge: Cambridge University Press.

Gilbert, Daniel T. 1991. How Mental Systems Believe. *American Psychologist* 46: 107–119.

Schiffer, Stephen. 1992. Boghossian on Externalism and Inference. *Philosophical Issues* 2: 29–38.

Spinoza, Baruch. 1985. *The Collected Works of Spinoza*, volume 1. Edward Curley, ed. and trans. Princeton: Princeton University Press.

Wegner, Daniel M. and James W. Pennebaker (eds.). 1993. *Handbook of Mental Control*. Englewood Cliffs, NJ: Prentice-Hall.