SANFORD GOLDBERG

EXTERNALISM AND AUTHORITATIVE KNOWLEDGE OF CONTENT: A NEW INCOMPATIBILIST STRATEGY

(Received in revised version in July 1998)

1. INTRODUCTION

A typical strategy of those who seek to show that externalism is compatible with authoritative knowledge of content is to show that externalism does nothing to undermine the claim that *all thinkers can at any time form correct and justified self-ascriptive judgements concerning their occurrent thoughts*. In reaction, most incompatibilists have assumed the burden of *denying* that externalism is compatible with this claim about self-ascription. Here I suggest another way to attack the compatibilist strategy. I aim to show that forming a justified true self-ascriptive judgement about one's occurrent thought *does not amount to or imply* that one 'knows the content' of the self-ascribed thought. While the difference between present-tense self-ascription and knowledge of content has previously been brought out using the familiar trappings of *world-switching* examples,[1] here I attempt to establish the difference by appeal to *actual (real-life) memory-involving* cases. To this end, I present a 'recollection problem' and argue that, so long as one conflates present-tense self-ascription and self-knowledge of content, there can be no satisfactory response to this problem. The result is that, even if the compatibilist strategy is correct in what it asserts about *self-ascription*, it has not delivered the relevant goods if it aims to establish a thesis asserting externalism's compatibility with *knowledge of content*. I conclude by speculating how the recollection argument to be presented here can be transformed, from an argument *against* the compatibilist strategy, into an argument *for* incompatibilism.

Philosophical Studies **100:** 51–79, 2000.

Externalism is the thesis that some contents cannot be individuated in terms of properties of the individual considered in isolation from her social and physical environment. The doctrine of authoritative self-knowledge of content is the view that, for any thinker *S* and occurrent thought that *p*, *S* can be said to have *a priori* knowledge of the content of her thought that *p*.[2] The central issue in this paper concerns the compatibility of these two doctrines. In denying this compatibility, incompatibilists are committed to the view that (given externalism[3]) there is at least one thinker who thinks a thought yet who fails to have *a priori* knowledge of the content of the thought she is thinking.[4] Since if *S* fails to know the content of her occurrent thought then (by extension) she fails to have *a priori* knowledge of the content of her thought, the incompatibilist need merely show that (given externalism) there is some thinker who thinks a thought yet fails to know its content.

In this paper I shall be taking aim at a doctrine that is central to the main strategy pursued by those seeking to vindicate the compatibilist position. The doctrine itself, 'SAKC' for 'self-ascriptive account of knowledge of content,' asserts that

> For all thinkers *S* and occurrent thoughts that *p*, if *S* forms the present-tense self-ascriptive judgement that she herself thinks that *p*, then *S* knows the content of the occurrent thought self-ascribed.

The role that SAKC plays in the debate concerning the compatibility of externalism and authoritative knowledge of content is to respond to a potential challenge that can be raised against Burge's original compatibilist position.[5] Burge argued (correctly, in my view) that thinkers' capacities to form correct, justified self-ascriptive judgements about their own occurrent thoughts derives from the self-referential character of those present-tense judgements. If correct, this would mean that no doctrine – and so by extension no externalist doctrine – can undermine a thinker's capacity to form correct, justified self-ascriptive judgements about her own occurrent thoughts. Incompatibilist worries arise, however, when we consider whether *correct, justified self-ascription of a content* can be conflated with *knowing the content one is presently self-ascribing*.[6] Suppose (what externalists appear ready

to acknowledge[7]) that knowledge of content requires knowing *what* content one is thinking; then such incompatibilist worries would appear to be exacerbated by the fact that most externalists concede, as an implication of their formulation of externalism itself, that

> (*) A thinker *S* may fail to be able to provide correct or exhaustive interpretations of the form of words *W* which *S* uses to express her occurrent thoughts, and *S* may fail to be able to discriminate her actual thought from other (counterfactual) thoughts she would have had if she had acquired her concepts in a different environment.

Now (*) may appear to jeopardize the doctrine of authoritative self-knowledge, and so lend ammunition to the incompatibilist, insofar as *failing to correctly interpret W* (or *failing to be able to discriminate the thought expressed by W from thoughts W would have expressed in other contexts*) might be taken as *failing to know what content one expressed with W*. This point would show that one can self-ascribe a thought with *W* yet fail to know the content of the thought one self-ascribed with those very words. It is in this context that SAKC plays its role in the central compatibilist strategy. Externalists who concede (*) as an implication of their version of externalism insist on SAKC as part of a strategy whose point is to deny that (*) opens up the possibility that *S* fails to know the content of her own thought.[8]

Incompatibilists have two options for responding to this compatibilist strategy. One is to challenge the idea that the ability of thinkers to form correct and justified self-ascriptive judgements about their occurrent thoughts derives from the self-referential nature of these present-tense judgements. While early incompatibilist arguments followed such a strategy,[9] I believe that there are many very good reasons to agree with the compatibilists on this point.[10] This leads me to the second (insufficiently appreciated) option open to the incompatibilist. This strategy would have the incompatibilist *grant* the point about present-tense self-ascription, but *deny* that present-tense self-ascription amounts to or otherwise entails knowledge of content. This is the option I have chosen.

In what follows I argue that SAKC faces a (heretofore unrecognized) difficulty, deriving not from the vagaries of world-switching,

but instead from facts about memory together with some well-entrenched belief-attribution practices. At the very least, my argument against SAKC will show that the case for compatibilism (based as it is on SAKC itself) is far from vindicated, with the result that the compatibilist owes us a new way to resist the inference that the incompatibilist would have us draw from (*). What is more, my argument arguably provides a new way to argue *for incompatibilism*; I will return to this in the last section of this paper.

2. THE RECOLLECTION PROBLEM

Let me begin, however, with the recollection argument against SAKC. Consider the following two examples:

> EXAMPLE 1. Judy, who has been out of college for quite some time, happens upon a paper that she had written for a philosophy class. On reading it she finds that it is unfamiliar to her. She thinks to herself: "What was I thinking?"
>
> EXAMPLE 2. Jones, a public official, is indignant over a recent episode in which a member of the press "misinterpreted" a comment he had made some time back. At a press conference, Jones is presented with a tape of the original interview. When pressed as to what he *really* meant, he found himself wondering about this (even as he knew that he *didn't* mean what the press reported him as having meant).

These examples identify and motivate the following intuition: when it comes to one's *past* thoughts, one may well be in a position at some later time to self-ascribe a thought using the *same form of words* as were used in the original expression of the thought, without thereby counting as knowing the content of the original thought itself. Consider: both Judy and Jones know the form of words *W* which they used to express their respective previous thoughts; presumably each can use *W* following a past-tense self-ascription such as, "I thought (meant) that . . . "; and yet even if they were to do so there would be no temptation to credit either one, merely in virtue of such a self-ascription, with self-knowledge of the content of the

thought self-ascribed. This shows that the restriction to '*present-tense*' judgements in SAKC is crucial: the self-ascriptive account of knowledge of content is false when the knowledge in question concerns one's *past* thoughts.[11]

Why *is* it natural to suppose that the self-ascriptive account fails to cover the case of knowing the content of one's own past thoughts? This question acquires added force when asked in the context of an endorsement of SAKC; for SAKC insists that, in *present*-tense self-ascriptions, the ability to use a form of words in self-ascribing a thought itself *suffices* to establish knowledge of the thought's content. Yet the examples above indicate that, when one makes a *past*-tense self-ascription, mere knowledge of the form of words in one's previously-produced autobiographical text is *not sufficient* for knowledge of the contents which this text was used to express. What accounts for this difference?

It is tempting to suppose that the whole difference between present- and past-tense self-ascriptions is one of *memory*. On this view, we do not credit Judy with knowledge of the contents of her past thoughts, even though she is in a position to self-ascribe a thought using the same form of words as she had used in the original expression of the past thought, merely because she does not *remember* what content she expressed with the text when she originally produced it.

This temptation, to explain the difference between present- and past-tense self-ascriptions in terms of a difference in *memory*, ought to be resisted. Memory issues *are* involved, to be sure; but reducing the issue to one of memory obscures the crucially important issue: *how can* one be in a position to misremember one's own thoughts (fail to remember their contents) when one knows the form of words which one had used to express those thoughts? What can such misremembering *consist in?* This question, which I will call the *recollection problem*, illuminates some important features of the relationship between self-ascription and knowledge of content – features that have been systematically neglected in the literature, owing to a limited focus on present-tense self-ascriptions.

That a narrow focus on *present*-tense cases distorts the relation between the relevant self-ascriptive statements (on the one

hand) and 'knowledge of content'-attributions (on the other), can be brought out in two steps.

First, consider what it is that Judy and Jones fail to know regarding their past thoughts. Given that they know the words they used to express their previous thoughts, the only thing that they can be plausibly construed as failing to know is *the content* of their past thoughts. In fact, we might even say that these examples help to fix *what it is that we're talking about* when we indulge in 'knowledge of content'-talk – we're talking about whatever it is that Judy and Jones fail to know in the memory cases just described.

Second, consider now that these examples fix the subject-matter of 'knowledge of content'-talk in a way that makes clear how *one can self-ascribe a content without counting as knowing the content self-ascribed*. This is an all-important virtue of the examples. In particular, the examples help to undermine what strikes me as a kind of theory-induced blindness to the difference between self-ascription and knowledge of content. The blindness in question is mainly evident among those externalists with compatibilist ambitions. The following conversation between a compatibilist and an incompatibilist is typical:

Compatibilist: I know that I'm thinking that water is wet.
Incompatibilist: But if externalism is true then you don't know the content of your thought; for, given externalism, you don't know whether you're thinking a water-thought or a twater-thought.
Compatibilist: Your point about my not knowing whether I'm thinking a water-thought or a twater-thought is, even if correct, simply irrelevant to the issue. I *do* know the content of my occurrent thought: I'm thinking *that water is wet*. This first-personal judgement of mine is both true (I'm thinking the very thought in self-ascribing it) and justified (because self-verifying as indicated). Thus, that I fail to be able to compare the thought I've just self-ascribed to other thoughts I've had (or might have had in some counterfactual world) is not relevant to the issue of whether I know the content of my present thought.[12]

Compatibilists appeal to this kind of reasoning to close off the *very possibility* that there might be more to knowing one's own occurrent contents than self-ascribing these contents. Now, while the examples of Judy and Jones do not establish any incompatibilist point by themselves – in fact, from what has been said so far they have nothing whatever to do with *present-tense* self-ascriptions – nonetheless these examples do give concrete sense to the idea that

one can make a *past-tense* self-ascription of a thought yet fail to know its content. This alone should suffice to prevent the sort of theory-induced blindness to which I have alluded.

At any rate, exchanges of the sort just described will be familiar to anyone following the literature on externalism and knowledge of content.[13] I for one have always had some sympathies for the incompatibilist side. At the same time I have been *most unsatisfied* with the way that incompatibilists have chosen to respond to the compatibilist counter-thrust. For most incompatibilists have tried to bring out their point using exotic Twin-Earth thought experiments.[14] But surely (I have thought) the point which incompatibilists seek to make *cannot* be so complicated that it requires such elaborate thought-experiment travel plans. This will be the real significance of the recollection problem: it presents what strikes me as the central incompatibilist insight – the difference between self-ascribing a content and knowing the content self-ascribed – using examples closely modeled on actual (as opposed to Twin-Earth-style counterfactual) memory-involving cases.

Having announced why I think that the recollection problem is novel, I now want to argue that, when it comes to attempting to answer this problem, anyone who endorses SAKC faces an insurmountable difficulty. Recall that the recollection problem is to analyze what it is to fail to know one's past thought (fail to 'know the content' of the past thought), despite knowing the form of words one used to express the thought. The difficulty is that an independently plausible assumption about what it is to express (and so self-ascribe) the same thought at two different times (call it 'STE' for 'Same Thought Expressed'), together with the assumption of SAKC, imply (contra our examples) that Judy's failure of knowledge of content *is impossible*. The difficulty facing proponents of SAKC, then, is that what appears to common sense to be a perfectly possible and even rather common phenomenon – that of failing to remember (and so failing to know) the content of one's past thoughts, despite knowing the form of words one used to express those thoughts – must be *impossible*. This result could be avoided if STE itself could be jettisoned in a plausible way; but STE cannot be jettisoned without undermining many of our most central belief-attribution practices. I treat this as a *reductio* of SAKC.

I shall proceed as follows. First, I will present and defend STE as stating a sufficient condition on expressing the same thought at two different times (section 2.1); I will suggest how SAKC and STE together imply that as described the memory cases of Judy and Jones should be impossible (section 2.2); and finally, I will consider and reject some ways in which proponents of SAKC might think to avoid this result (section 2.3). Having done so, I will offer some programmatic remarks about how the recollection problem might be handled, once we surrender SAKC (section 3). In doing so I aim to suggest that the present *reductio* is not just an argument against SAKC, but against compatibilist ambitions more generally.

2.1. *STE ('Same Thought Expressed')*

Here I begin by stating a principle meant to be a sufficient (but not necessary) condition on expressing the same thought at two different times. The principle is this:

> (STE) A single individual *S* expresses the *same thought that p* at t_2 as the thought she expressed at (some earlier time) t_1 when (i) she uses the same form of words *W* at t_2 as those she had used at t_1 to express the thought, and *W* is part of the lexicon of a language which *S* understands at both t_1 and t_2 (on any acceptable notion of what it is to 'understand a language'[15]); (ii) *S* intends at t_2 that *W* means the same for her now (at t_2) as it did at t_1; and (iii) the statement made by *S*'s use of *W* at t_2 and the statement made by *S*'s use of *W* at t_1 have the same truth-conditions.[16]

My thesis shall be that STE is an acceptable *sufficient condition* on expressing the same thought at two different times. Any principle *weaker* (less demanding) than STE is unacceptable, as it will yield results which would have us hold that a thinker expresses *the same thought* at two different times, in situations in which standard belief-attribution practices would have us treat the utterances as expressing *distinct* thoughts. And anything *stronger* (more demanding) than STE is unacceptable, as it will provide results which would have us hold that a thinker expresses *two distinct thoughts* on two different occasions, in situations in which standard belief-attribution prac-

tices would have us treat the utterances as expressing the *same* thought. Thus my case for STE consists in showing that it alone coheres with standard belief-attribution practices. As befits this sort of argument, I will bring this out by employing the method of reflective equilibrium: examining intuitively clear cases in which it is easy to reach a verdict whether a thinker's two utterances express the same thought, reflecting on the principles which might account for our intuitive verdicts – and adjusting both sides where properly motivated.

Return to Jones' case above. Suppose that the sentence whose interpretation is in question is

(J) "I believe that public officials go beyond their legitimate authority when they make decisions which do not reflect the actual desires of a majority of their constituents."

Now consider the following elaboration. Jones uttered this sentence at the original interview. Subsequently, he was interpreted as meaning that public officials who make such decisions do something which is *illegal*. He denied that this was what he meant. When presented with a tape on which he heard himself utter (J) and queried as to what he meant, he responds by repeating *the exact words* he used in (J) itself. (We suppose that he did so in order to buy some time.) Under such circumstances it would be most intuitive to describe Jones as 'saying the same thing' (and so as 'expressing the same thought'[17]) but in an uninformative manner. This much is patent. But if we accept this verdict then we are committed to the idea that Jones can express the same thought – that he can 'say the same thing' – even when he himself acknowledges that he does not know exactly what it is he is saying (that is, when he cannot explicate the concepts that figure in the thought in a way that even he would find satisfactory).

In fact, it is easy to see that accepting this intuitive verdict (that Jones is to be described as having 'said the same thing') will commit us to more. First, it will commit us to the idea that decisions regarding whether a person has 'said the same thing' on two different occasions are sometimes settled simply in virtue of the person's avowed semantic intention to have said the same thing. Consider: the reason *why* we describe Jones as 'saying the same thing' despite his not being able to explicate exactly what it is that

he is saying, is that *he intends* that his words mean the same thing. This is a point about the kind evidence that would convince us that he is to be treated as having said the same thing: his *avowing* as much would be sufficient. However, since one's semantic intentions to 'say the same thing' might be defeated – below I will identify two ways – we should conclude only that one's semantic intentions enjoy the *presumption of truth*.

But second, accepting the intuitive verdict that Jones has in fact 'said the same thing' will commit us to a particular conception of *what it is* to intend to say the same thing. In particular, it will commit us to holding that one satisfies the *same-intended-meaning-condition* (condition (ii) of STE) if one can form the relevant indexically-specified intention. (I will call this the *indexical intention assumption*.) That accepting this intuitive verdict in Jones' case commits us to the indexical intention assumption can be brought out as follows. Our first point was that Jones is treated as 'saying the same thing' on the two distinct occasions *because* he avowed as much on the latter occasion. This assumes that in fact he has formed the requisite semantic intention to say (or mean) the same thing. Yet if the conditions for satisfying the *same-intended-meaning* condition require anything more than that Jones be able to form the relevant indexically-specified intention, then our description of Jones (as 'saying the same thing') will be in jeopardy. For, given any proposed condition on *meaning the same thing with the same form of words* which is *more stringent* than the indexical intention assumption, it is possible that Jones' fails the more stringent conditions on intending to mean the same thing. When he does, he will not count as having the requisite semantic intention to say (or mean)[18] the same thing. But if he does not count as having the requisite semantic intention to say (mean) the same thing, then it appears that we will have no reason to describe Jones as 'saying the same thing' (since the having of that intention was what warranted this description). But we *do* have a reason to describe Jones as 'saying the same thing' (namely, his avowed intention). Thus we reach a conception of what is sufficient for having the intention to say the same thing, namely, the indexical intention assumption.

Nor is Jones' case the only type of case where a more stringent account of *intending to mean the same thing* will run up against

one of our well-entrenched intentionality-ascription practices. Consider the phenomenon whereby we describe ourselves and others as 'believing what another person says.' In many cases, when one person *S* comes to believe what another person *T* says, *S* also acquires the disposition to *express* the belief she acquires with the *same form of words* as those *T* used. But suppose (what is often the case) that *S*'s grasp of the meaning of those words is less than complete; then it may be that the only recourse that *S* has to fixing the meaning of the words is to form the requisite indexically-specified intention. So for example suppose that Jen, who knows nothing about astronomy, hears her professor utter the sentence, "There is a black hole in Galaxy *X*," and that on this basis Jen acquires a belief she would express with that same sentence. It is natural to suppose that the belief she expresses is the belief that there is a black hole in Galaxy *X*; yet, if Jen is completely ignorant of astronomy, then upon being queried as to what she means she may respond by saying that she means (intends to mean) exactly what her professor meant (*whatever he meant*). Were it not for the fact that her mere indexically-specified intention sufficed as an intention to 'say the same thing' as her professor, she would not count as having the intention to say the same thing – with the unacceptable result that we would have no reason to treat her as 'believing what her professor said.'

Here, as in the case of Jones, we see that there are cases of belief-attribution in which the intuitive verdict requires endorsing the indexical intention assumption. But *what is it* to form the requisite indexically-specified semantic intention? Our assumption has been that

> A sufficient condition for a person *S*'s intending her words to have the same meanings at t_2 as they had for her at some previous time t_1 is that *S* has the appropriate *indexically-specified* intention.

The view in question is that, to the extent that a thinker *can refer* at t_2 to a previous occasion (say, at t_1) on which the same form of words *W* were used as she is using at t_2, then she is in a position at t_2 to intend that her words mean what they meant *then* (at t_1), and can do so merely by forming the appropriate indexically-specified

intention. The 'can refer' in the antecedent is important. If a thinker has used *W* on many previous occasions, and has at present no way to single out one particular occasion of use – say, all previous occasions are a blur to her – then she is not in a position to be able to refer to *one particular* occasion of use. Thus we stipulate, as part of the indexical intention assumption, that a thinker's intention is not of the 'appropriate' sort to refer to a particular previous occasion unless she has some *independent* way – some way that is independent of her present use of indexicals such as 'then' – to single out a previous occasion of use. That a thinker *S* can satisfy this and still form what can be labeled an 'indexically-specified' intention can be made clear by example: *S* might intend that her words mean now *whatever they meant* when she used them at 7:35 p.m. last Saturday (August 1, 1998) in Café Phoenix.

Now it might be thought that there is an acceptable version of STE which is weaker than the version I gave above. On this view, it is sufficient for a person to express the same thought at two distinct times if she satisfies conditions (i) and (ii) of STE. While it would help my case against SAKC immensely if this weaker version of STE were acceptable, unfortunately it is not. In fact, there are (at least) two kinds of example in which a thinker *S* satisfies (i) and (ii) of STE, yet (arguably) fails to express the same thought as her earlier thought. Schematically these can be understood as cases in which *S*'s semantic intention *is defeated* by other considerations, as for example when (1) a thinker possesses a suitably large number of false beliefs about the subject matter of the words towards which she is directing her indexically-specified intention, or else (2) there is a change in some external condition relevant to individuating the meaning of *S*'s words, which change goes unnoticed by *S* (i.e., Twin-Earth-type considerations).

Consider the following illustration of the way in which considerations of type (1) might undermine a thinker's semantic intentions to 'say the same thing'.[19] Once, as a small child, Mathilde spent a lovely autumn in the pretty parts of London, England. On the occasion of her seventh birthday (August 1, 1918) she uttered these words to her mother:

(M) I think that London is pretty.

That was the last time that Mathilde uttered that sentence. She has since forgotten everything about her time in London except the fact that she uttered (M) to her mother. She does not remember *where* she uttered those words but she does remember *when* (the memorable occasion of her seventh birthday, which just happened to have been the last birthday that she would celebrate with her mother). In 1998 Mathilde now believes (falsely) that she uttered those words in London, Ontario, where she has resided for the past eighty years. She also believes (again, falsely) that she has never been to London, England. Pictures and other mementoes of her trip have either been destroyed or lost. Family and friends who could testify about her time in England are no longer living. She has absolutely no recollection of ever having been in England. On August 1, 1998, an old friend from the States visits Mathilde in honor of her eighty-seventh birthday. Her friend asks her what she thinks of London (meaning London, *Ontario*). This brings to Mathilde's mind thoughts of her long-deceased mother, which prompt her (Mathilde) to think to herself, "Hmmm, I'll tell my friend exactly what I told my mother years ago …" – following which Mathilde utters (M) with the avowed intention that the words "London is pretty" mean the same for her now as they did back on August 1, 1918. Only given her memory lapse, Mathilde now thinks that, on both occasions, she meant by her utterance the same thing (namely, that London, Ontario – the city in which she has lived for most of her life – is pretty).

Notice that Mathilde satisfies conditions (i) and (ii) of STE, yet for all that there is little temptation to construe her as expressing the same thought when she utters (M) now as the thought she expressed when she uttered (M) in 1918. The truth-conditions of her utterance of (M) in 1918 are that London, England is pretty; the truth-conditions of her utterance of (M) in 1998 are that London, Ontario is pretty. To substantiate this we might appeal to the idea that London Ontario is the dominant source of information[20] which Mathilde associates with 'London.' This example brings out the point of condition (iii) of STE: a thinker who satisfies (i) and (ii) of STE still may fail to express the same thought as an earlier thought she expressed, on the grounds that her semantic intentions are under-

mined by considerations which suggest that the statements that she made with these words on these two occasions differ in truth-value.

Nor is the case involving a suitable number of false beliefs about one's subject-matter the only case in which one's semantic intentions can be undermined. There are examples of type (2) which can also (be thought to) undermine a thinker's intention to 'say the same thing.' Here the point is that an externalism about content might be used to show that one's words can fail to mean what they meant previously, for reasons which have to do with changes in the nature of the external (social and physical) context in which one acquires and uses the words. Consider: an Earthian and a Twin-Earthian who both utter the form of words 'Water is wet' say two different things (i.e., express statements that differ in regard to their truth-conditions). Now suppose that, unbeknownst to her, *S* is world-switched to Twin-Earth and stays long enough to acquire the concept *twater*. Then it is plausible that *S*'s utterances of 'Water is wet' at the two distinct times express different thoughts, *despite S*'s avowed intention at the later time to have said the same thing with these words as she had said at the earlier time.[21] Thus it is clear how externalist considerations might be used to show how a thinker's avowed semantic intention to 'say the same thing' can be undermined.

The Mathilde and world-switching examples show that if STE aims to capture a sufficient condition for expressing the same thought at two distinct times, and seeks to do so in a way that coheres with our standard belief-attribution practices, it must go beyond conditions specifying same words, of a still-understood language, uttered with the semantic intention to 'say the same thing.' STE must include a condition stipulating that the words *S* uses (with the appropriate semantic intention) must have *the same truth-conditional meaning* on the two occasions.[22]

I submit that STE, formulated in terms of (i)–(iii), amounts to a sufficient condition on expressing the same thought at two different times. I will defend this claim in greater detail below (2.3), when I consider how a proponent of SAKC might modify STE in order to avoid counterintuitive results in the memory case. I begin, however, by trying to bring out those results.

2.2. *STE and SAKC Imply the Impossibility of the Memory Cases*

I will now proceed to show that, on the assumption of STE and SAKC, Judy and Jones *do* know the content of their past thoughts (our intuition to the contrary notwithstanding). Suppose some thinker *S* satisfies (i)–(iii) of STE. It would follow that, by uttering at t_2 the words *W* which she had previously used in the expression of her thought at t_1, *S* would count at t_2 as expressing the *same thought* as the thought she expressed at t_1. What is more, given SAKC, if *S* were to use *W* in the context of a *present*-tense self-ascription at t_2, *S* would *ipso facto* count as *knowing the content* of the thought presently self-ascribed. But if she *knows the content* of her occurrent thought, and if the occurrent thought is *the same thought* as the past thought, then she knows the content of her past thought! The problem with this result is clear: our example of Judy above supports the idea that it is *deeply counterintuitive* to count Judy as knowing the content of her (previous) thought merely in virtue of being able to make a present-tense self-ascription with the same form of words as she had used earlier: even if she intends these words to mean what they meant earlier, it seems clear that she does not know what this meaning is!

This argument has the form of a *reductio*: two seemingly plausible premises (STE and SAKC) lead to an unacceptable conclusion (the impossibility of failing to remember the content of one's past thought in the cases described). It will be clear what premise I think must go. Given the facts about memory and the plausibility of STE, we should abandon SAKC (as having unacceptable implications regarding memory cases). I have little doubt that most compatibilist externalists will not follow me in this. I will now argue, however, that none of the avenues for resisting this conclusion are attractive.

2.3. *How Not to Respond to the Argument So Far*

In this section I will consider and reject three distinct ways in which the externalist proponent of SAKC might think to respond to the argument so far. All three responses are based on one and the same strategy. On this strategy, the proponent of SAKC *denies* that Judy is self-ascribing the *same thought* now as she had expressed at that earlier time. For, having challenged this idea, such a proponent could *allow* that Judy knows the content of the thought she is

presently self-ascribing, without implying (what the example of Judy suggests is an absurd claim) that she knows the content of the earlier thought she had expressed with those same words.

There would appear to be only three ways in which an externalist can justify the *different thoughts thesis* (the thesis that Judy expresses a different thought-content at t_2 than she had expressed at t_1): by arguing that STE does not amount to a sufficient condition on expressing the same thought at two different times; by arguing that condition (ii) is not satisfied in Judy's case; or by arguing that condition (iii) is not satisfied in Judy's case. None of these ways is attractive.

2.3.1. *Why We Should Not Reject STE*

Let us first examine the attempt to react to the *reductio* by rejecting STE. By now it should be evident that such a move comes at too high a cost. For if the fact that Judy satisfies conditions (i)–(iii) of STE does not itself warrant the claim that she expresses the same thought at $t_2$as the thought she expressed at t_1, then the same holds true in the extended case of Jones. We recall that, when asked as to what he meant by (J), he says, "I meant that . . . ," where in the place of the ". . . " he repeats *the very same words* (with the appropriate indexically-specified intention). But we saw that it was implausible in the extreme (and so perfectly *ad hoc*) to count Jones as saying something different at the time of recollection from what he said at the time of the original interview. In short, the move to abandon STE has this unacceptable implication in Jones' case, and so rejecting STE is not a plausible way to react to the *reductio*.

In fact, it would appear that we can make a second, even stronger point against the proposal to reject STE itself. If Judy does *not* express the same thought at t_2 as the thought she had expressed att_1 despite satisfying STE's (i)–(iii), then we are threatened with the possibility that Judy will never be able to *recall* her previous thought. For, if Judy's satisfying (i)–(iii) is insufficient to construe Judy as re-expressing an earlier thought, then it seems as though *nothing* will suffice to construe her as expressing that previous thought *at all*. (To see this, let us ask how she might recoup the ability to express the previous thought; it would appear that the answer is that, either she does so by satisfying STE itself, or else

– for all practical purposes – she cannot do so at all.) But if she cannot re-express the thought at all, she can no longer *self-ascribe* it (and so cannot *recall* it) at all. In short, the present reaction to the difficulty, which started off in the hope of saving SAKC by denying STE, threatens to degenerate into a form of skepticism concerning the very possibility of being able to recall past thoughts. And if it does degenerate, then this result is plausibly regarded as a *reductio*: the skeptical conclusion follows, not from any straightforward memory failure, but from a doctrine about the relation between self-ascription and knowledge of content.

2.3.2. *Why We should Not Reject that Judy Satisfies Condition (ii) of STE*

Let us now consider a second way in which a proponent of SAKC might think to avoid the argument so far presented. On this way, we leave STE in place, and reject instead the idea that Judy satisfies condition (ii) of STE. Since this move is tantamount to denying the indexical intention assumption, most of my objections will have been anticipated; I have three.

First, denying that Judy satisfies condition (ii) will entail an overly-rigorous conception of what counts as a thinker's expressing the same thought on two different occasions; this was demonstrated in 2.1, where we saw that any view that denies that Judy satisfies (ii) flies in the face of well-entrenched belief-attribution practices.

Second, rejecting (or otherwise modifying) the indexical intention assumption threatens to land one with the same absurdity as the absurdity that appears to be implied by the proposal to reject STE. For, to the extent that we increase the requirements on satisfying the *same-intended-meaning* condition, we make it increasingly unlikely that *S* will be in a position to *intend* to express the same meaning with the same form of words at two different times – with the result that we make it increasingly unlikely that *S* will be in a position to *mean the same thing* with the same form of words at two different times. Since this is just what prompted our second criticism of the proposal to reject STE, the same criticism as above is in place here.

But there is a third reason in support of the indexical intention assumption. Consider: *the very considerations* which lead Burge and others to hold (plausibly) that a thinker can have a thought

without knowing the conditions that individuate that thought, should lead us to hold that a thinker can form an intention without knowing the conditions that individuate that intention (for any intention whatever). For both cases should be covered by the same Burgean point: a sufficient condition on a thinker's having a mental state with content that *p* is that she have at her disposal a form of words whose utterance in the assertive mode expresses the content that *p*. Burge has asserted (most plausibly) that this holds when the mental state in question is the belief that water is wet; it should also hold when the mental state in question is the intention that one's words mean now what they meant at some earlier time. But it seems that an indexically-specified utterance of the proper sort should *suffice* for expressing such an intention.

To conclude. The indexical intention assumption is reasonable on grounds independent of the present considerations, and so a proponent of SAKC cannot plausibly deny this assumption. But if the proponent of SAKC cannot plausibly deny this assumption, then the proponent of SAKC cannot plausibly maintain that Judy fails to satisfy condition (ii) of STE.

2.3.3. *Why We Should Not Reject that Judy Satisfies Condition (iii) of STE*

The third and final response to be considered here is this: the *reductio* argument of section 2.2 fails because in the cases of Judy and Jones their words *do* in fact change their truth-conditional meanings; that is, because condition (iii) is not satisfied.

This move seems *ad hoc* right from the start. For, if at some time t_2 a thinker avows an intention to the effect that her use of words *W* at t_2 means what she meant with *W* at t_1 (some earlier time), then surely the burden of proof is on those who would assert that the statement made by her use of *W* at t_2 has different truth-conditions than the truth-conditions of the statement made by her use of *W* at t_1. The thinker's intentions do not *settle* the matter whether two of her statements differ in truth-conditions; but they do enjoy the benefit of the doubt. With this presumption in place, Judy's case clearly suggests how *ad hoc* it would be to say that her words change in truth-conditional meaning. For it is clear that her

case cannot be assimilated into either of two ways in which one's semantic intentions might legitimately be said to be undermined.

Consider first way (1), according to which a thinker's semantic intentions to 'say the same thing' are undermined because she has some suitably high number of false beliefs about the subject-matter of the words on which she is fixing her indexically-specified semantic intention. (We illustrated this with the case of Mathilde.) Let us now return to the case of Judy, who (by her own admission) does not know *now* what she was thinking when she wrote her philosophy paper some time ago.[23] And let us ask: is what Judy does not remember regarding the content of her past thought sufficient to show that when she uses the same words *now* (time of recollection) as those she used *then* (time of original paper), these words change in truth-conditional meaning? The question, of course, is this: what counts as a 'suitable number' of false surrounding beliefs, such that these false surrounding beliefs undermine a person's semantic intention to use her words to 'say the same thing' as she said before?

To answer this, consider what it was in the case of Mathilde that undermined her semantic intentions. As I see it, three factors were involved: Mathilde's life-long residency in London *Ontario*; the false empirical beliefs she presently has concerning her whereabouts when she uttered (M) in 1918; and her failure to remember her one-time visit to London *England*. Together, these three factors make the following claim plausible: even though Mathilde *claims* that her present utterance of 'London is pretty' is intended to mean what she meant when she uttered this string of words to her mother in 1918, nonetheless the truth-conditions of the two statements she made by uttering (M) – the one in 1918, the other in 1998 – differ.

But Judy's case is different in the relevant respects. For while Mathilde had all sorts of memory problems and false beliefs about where she was when she originally uttered (M) to her mother in 1918, Judy's memory failure is more limited. Let us suppose that Judy remembers the professor of the course, the course themes (in broad outline), and the authors that were read. And let us suppose that what she forgets is restricted: she forgets what she meant with a number of the sentences that figure in the paper she wrote. To be sure, this kind of forgetfulness involves forgetting the main inferential connections of the concepts she was using and the exact context

of the issues she was treating. But is what she forgot sufficient to undermine her intention to use the words to say the same thing as she had said earlier (i.e., when she wrote the paper)?

Once again, any temptation we might have in the direction of an affirmative answer dissipates once we realize that we are going to have to treat Jones' case in an identical fashion. Thus, if we say that Judy's semantic intentions are undermined because her words change in their truth-conditional meaning, then we will have to say that Jones' semantic intentions are undermined for the same reason; and yet in the latter case this claim was revealed to be deeply offensive to our intuitions (and so *ad hoc*). The result is that Judy's case is relevantly different from Mathilde's case: simply put, absent quite a bit of forgetfulness relevant to the subject-matter of the utterances themselves, the claim that a thinker's semantic intentions are undermined by what she forgets does not appear to be justified in the least.

Turn to the proposal to treat condition (iii) as unmet in Judy's case, on grounds deriving from an externalism about content. Since Judy's failure of memory is a failure of memory *right here* on earth, and remains so even when it is *stipulated* that she remains in the same linguistic community all along, and has no doubts on this score, this response is a non-starter.

In short, neither of the two ways of denying that Judy satisfies (iii) appear plausible. More generally, we have seen that none of the various ways of pursuing the different thoughts thesis – rejecting STE, rejecting that condition (ii) is satisfied, rejecting that condition (iii) is satisfied – can avoid the *reductio* argument of 2.2. To be sure, one might try to figure out how to establish the different thoughts thesis without running into the looming *reductio* argument. I for one don't see how this can be done, consistent with not offending against one or another well-entrenched belief-attribution principle. But, since I may be overlooking some possibility, I will put the point of the preceding in the form of a challenge posed by the recollection problem: on the assumption of SAKC, how can we explain what it is to fail to know what one oneself was thinking, when that failure of knowledge occurs *despite* knowing the form of words used to express the past thought? This is a 'problem' for those who endorse SAKC precisely because an answer appears to require a distinc-

tion between *expressing* (or *self-ascribing*) a thought and *knowing the content* of the thought self-ascribed – yet such a distinction undermines SAKC itself.

3. THE STATUS OF THE ARGUMENT SO FAR

So far I claim to have shown that, as a conception of what it is to know one's own occurrent contents, SAKC is objectionable. The upshot of the forgoing is that, if in the recollection cases Judy does in fact make a present-tense self-ascription of a thought in a way that satisfies (i)–(iii) of STE, then she would be making a present-tense self-ascription of a thought, yet would not count as knowing the content of the thought thereby self-ascribed.

This result is relevant to the compatibilism debate by establishing that, insofar as it relies on the strategy of insisting on SAKC, the case for compatibilism has yet to be vindicated. We recall that incompatibilists cite considerations such as (*) as part of their challenge to compatibilists to show how it is that, given externalism, a thinker *cannot but* know the content of the thought she is thinking. We recall further that in response the central compatibilist strategy has been to insist that it is sufficient, in order to count as knowing the content of the thought one oneself is thinking, that one be able to make a present-tense self-ascription of the thought in question.[24] What the recollection argument shows is that this implicit appeal to SAKC itself is illegitimate. Thus, the recollection argument helps the incompatibilist side by serving as a challenge to the compatibilist: given that SAKC can be seen on independent grounds to be inadequate as a conception of knowledge of one's own contents, why should we accept SAKC in cases where what is in question is the compatibility of externalism and knowledge of content?

But now it is a noteworthy feature of the way in which I criticized SAKC that my argument had *nothing whatever* to do with externalist considerations. In light of this it might be thought that, if the argument succeeds, it succeeds at showing that there is a problem of knowledge of content for *anyon*e, *regardless* of their commitments on the score of externalism. Actually, I have great sympathy for this claim.[25] At the same time, I believe that the recollection argument can in fact be used to establish points against externalism in partic-

ular. We can see this by considering just how we should answer the recollection problem itself.

4. SOME PROGRAMMATIC REFLECTIONS ON THE RECOLLECTION PROBLEM

Having formulated the recollection problem and having stated why I think it poses a problem for proponents of SAKC, I now wish to indicate what form an answer to the recollection problem might take, once we surrender SAKC. It turns out that the best answer to this problem makes it much easier to show how externalism jeopardizes authoritative knowledge of content. In a previous article[26] I formulated a principle asserting what does *not* count as knowingly identifying one's own thought-contents, in the form of the Principle of Knowing Identification:

(PKI) If *S* self-ascribes a thought with a form of words *W* which is such that,

(i) by *S*'s own lights, there is more than one interpretation that can be attached to *W*, *and*

(ii) *S* herself has at present no illuminating way to specify one over the other as the interpretation she intended,

then *S*'s self-ascription does not count as self-knowledge – because it is not a knowledgeable identification – of the content of the thought in question.

Since any *S* who satisfies PKI's two conditions *ipso facto* can be described as not having 'knowingly identified' the content of her thought despite having known the words to express the thought, we might try to answer the recollection problem by seeing what it takes to *avoid* satisfying PKI. This shall be my strategy.

On my view, *S* does not satisfy PKI so long as she can produce an *identifying interpretation* of *W*. As a first approximation, *S*'s interpretation of her own words *W* is an identifying interpretation when, given some third party *Q* who has certain questions concerning the proper interpretation of *W*, *S*'s own interpretation can answer these questions to *Q*'s (rational) satisfaction. In line with this, the knowl-

edge failures involved in the recollection problem can be described as a limiting case, in which the thinker fails *by her own lights* to produce such an interpretation. Let me explain.

Take the case of Jones, who holds that he used 'legitimate authority' to mean something other than what the journalist had attributed to him. In such a context, there are all sorts of practical issues that a disinterested third party *Q* might have in mind: Why did he (Jones) choose the words 'legitimate authority'? What belief did he want to convey to his audience? Was he aware of this possibility of misinterpretation based on the contrast between 'legitimate authority' and 'illegal'? etc. I submit that, as alternative interpretations of his previously-produced autobiographical text become more and more plausible – that is, as interpretations *other than his own interpretation* become more and more plausible – these interpretations serve as a benchmark against which *Q* can assess Jones' attempt to vindicate himself on the score of knowing the content of his own (previous) thoughts. In particular, such alternative interpretations provide a group of candidate interpretations, against which Jones' own self-interpretation can then be evaluated as more or less plausible. To be sure, in such cases there is still be the general presumption that Jones' own interpretations are authoritative (even adjusting for memory failure). But this authority is not absolute: in any particular case the presumption of authority might be overturned if it turns out that Jones' own interpretations are less plausible than the alternatives.

In line with this, we can now explain Jones' own recollection failure as follows. In 'not knowing (at t_2) what he was thinking (at t_1)' Jones himself plays a role which in one respect is like that of *Q*: Jones is aware that his own present interpretation of his own previously-used words does not provide good answers to certain practical questions that might arise about his own previously-produced autobiographical text. Unlike *Q*, however, Jones presumably does not have access to other candidate interpretations which would do a better job in this regard. I submit that this analysis, whereby *S* has self-knowledge of the content of her thought only if she can provide an identifying interpretation of the form of words she uses to express her thought, provides a natural explanation of how it is possible for *S* to fail to know her past content despite

knowing the form of words that she originally used to express that content.

No doubt, compatibilist externalists will resist this programmatic analysis, since the answer makes the case for incompatibilism that much easier. As I said in section 1, most proponents of externalism grant, as an implication of their version of externalism, that a thinker *S* may fail to be able to provide a correct or exhaustive interpretation of her own words, or may not be able to discriminate her present thought from other content-distinct thoughts. In light of this, given the requirement of an identifying interpretation, it would seem that incompatibilism would be vindicated so long as there is a case in which a third party *Q* is on the scene whose interest in querying *S*'s knowledge of her content derives from interests which *S*'s own answers do not settle. But, if the motivations prompting a compatibilist externalist to reject the programmatic analysis are clear, the analysis itself derives support from my argument in section 2. In particular, externalists who would reject this analysis, and would do so by appeal to SAKC, face the unmet challenge of having to answer the recollection problem itself.

I do not pretend that the case for incompatibilism presented here is anything other than horribly brief and sketchy, and I acknowledge that much more needs to be said before I can claim to have presented a compelling argument for incompatibilism. But, in order to generate some enthusiasm for such a project, I would like to conclude by emphasizing the uniqueness of my would-be incompatibilist argument from recollection. In the form suggested here, the incompatibilist argument is advanced in two stages. First, it is argued that the best answer to the recollection problem involves appeal to the notion of an identifying interpretation. Second, it is argued that the call for an identifying interpretation, together with any version of externalism that acknowledges (*) as an implication, will yield the result that (so long as some properly-situated *Q* is on the scene, equipped with queries which the thinker herself cannot answer) such a thinker fails to know the content of her own occurrent thoughts. If this sort of argument for incompatibilism should prove acceptable, it would be unlike almost every other incompatibilist argument with which I am familiar,[27] for the simple reason that it does not advert to counterfactual world-switching

cases, or even to more generic skeptical considerations. Rather, it would seek to make its point by appeal to memory-involving cases drawn from real life, and would do so by having previously motivated a novel (context-sensitive) conception of 'knowledge of content'-attributions.

In any case, whether it is used as part of an argument *against the central compatibilist strategy* or as part of an argument *for incompatibilism*, the recollection problem supports an idea that I for one have suspected all along. This is the idea that recent attempts to reconcile externalism with authoritative knowledge of content suffer from a problem that is at once *more basic* and *less philosophical* that one would expect, given the fanciful kinds of arguments traditionally advanced on behalf of incompatibilism. One need not travel to Twin Earth to see the problems with compatibilist positions. One need go no further than a careful consideration of what is involved in past-tense self-ascriptions, and how self-ascriptions in general are related to 'knowledge of content'-attributions. To see this, we can remain right here on Earth.

ACKNOWLEDGEMENTS

I would like to thank Akeel Bilgrami, Tony Brueckner, Joe Cummins, Ray Elugardo, Sidney Morgenbesser, Richard Muller, and John Stone for their helpful comments on earlier versions of this paper; and those who attended the session of the 1997 Central States Philosophical Association meeting, where I presented this paper. (Special thanks to Ray Elugardo for being such a helpful commentator at that session.)

NOTES

[1] See Sanford Goldberg, "Self-Ascription, Self-Knowledge, and the Memory Argument," *Analysis* 57, No. 3 (1997); and Paul Boghossian, "Content and Self-Knowledge," *Philosophical Topics* 17, No. 1 (1989).

[2] Nothing will hang on my use of '*a priori*'; for present purposes '*a priori* knowledge' is roughly equivalent to 'knowledge whose justification is not dependent on empirical investigation.'

[3] As will emerge below, the claim holds for any version of externalism which

allows that a thinker may have incorrect or otherwise incomplete explicational knowledge of the concepts that figure in her thoughts. Since most versions of externalism *do* allow for this – especially the popular ones, such as Putnam's and Burge's – I will not bother to qualify 'externalism.' But this qualification ought to be kept in mind, since Bilgrami's "Can Externalism be Reconciled with Self-Knowledge?" *Philosophical Topics* 20, No. 1 (1992) has made the case for a version of externalism that does not allow for incomplete explicational knowledge. (I should add that I do not endorse Bilgrami's view, but for reasons which are not relevant to the present themes.)

[4] This manner of setting up what is at issue in the compatibilism debate is not universally accepted. For example, in "Externalism, privileged self-knowledge, and the irrelevance of slow switching," *Analysis* 57, No. 4. (1997), Ted Warfield argues that the case for incompatibilism requires much more than showing what I claim the incompatibilist must show. However, I believe that if one reads the seminal article on this issue – Burge's "Individualism and Self-Knowledge," *Journal of Philosophy* 85 (1988) – one sees that Burge was interested in securing the compatibility of his anti-individualist externalism with the universally-quantified sentence that, for all thinkers *S* and thoughts that *p*, if *S* thinks the thought that *p* then *S* has 'basic self-knowledge' of the thought that *p* (note the implicit universal quantifier in his comments on the 'self-verifying' judgements involved in 'basic self-knowledge,' p. 649). Thus, if Burge's argument was meant to be an argument for the compatibility of externalism and authoritative *knowledge of content* (as distinct from 'basic self-knowledge'), then I am correct to set up the compatibility issue as I have.

[5] See Burge (1988). For variations on this theme, see John Gibbons, "Externalism and Knowledge of Content," *The Philosophical Review* 105, No. 3 (1996); and Rockney Jacobsen, "Self-Quotation and Self-Knowledge," *Synthese* 110 (1997).

[6] It might be wondered what, beyond the ability to make correct and justified (present-tense) self-ascriptions of one's thought, is involved in knowledge of one's own occurrent contents. Tony Brueckner raises just this issue, in his response to my (1997); see his "Difficulties in Generating Skepticism about Knowledge of Content," *Analysis* (forthcoming). But the novelty of my attempts to generate an incompatibilist view, both in my (1997) and here, *just is* that they do so without challenging anything that compatibilists have claimed regarding the truth and justification of present-tense self-ascriptions. I present my recollection-based argument here in sections 2–3; and in section 4 I reflect on this argument to suggest what else is presupposed by attributions of self-knowledge of content to a thinker.

[7] See Burge (1988) p. 662.

[8] For examples of this use of SAKC by externalists with compatibilist ambitions, see Burge (1988); Burge, "Wherein is Language Social?" in George, ed., *Reflections on Chomsky* (Oxford: Basil Blackwell, 1989); Falvey and Owens, "Externalism, Self-Knowledge, and Skepticism," *The Philosophical Review* 103, No. 1 (1994), p. 123; and Gibbons (1996), p. 302.

[9] See Tony Brueckner, "Skepticism of Knowledge of Content," *Mind* 99 (1990); and arguably Boghossian (1989).

[10] See Burge (1988), Falvey and Owens (1994), Gibbons (1996), and Jacobsen (1997).

[11] This claim is defended throughout section 2. In any case a similar point is made, in connection with the doctrine of externalism, in Boghossian (1989). However, Boghossian's manner of making this point leaves it open to some objections; see Peter Ludlow, "Social Externalism, Self-Knowledge, and Memory," *Analysis* 55, No. 3 (1995b); and Goldberg (1997).

[12] The compatibilist rejoinder here is modeled on Falvey and Owens (1994).

[13] The main source for this kind of compatibilist move is found in Burge (1988), John Heil "Privileged Access," *Mind* 97, No. 386 (1988); Falvey and Owens (1994); and Gibbons (1996).

[14] I myself am guilty of having made such an argument; see Goldberg (1997). In any case not all incompatibilist arguments seek to make their point by using Twin-Earth examples. For examples of incompatibilist arguments that do not require the stage-setting of Twin Earth, see Bilgrami (1992) (where an incompatibilist argument against Burge's anti-individualism is pressed into service for Bilgrami's version of non-social externalism) and Ludlow "Externalism, Self-Knowledge, and the Prevalence of Slow Switching," *Analysis* 55, No. 1 (1995a), where he examines an argument that he himself does not accept (for which see Ludlow (1995b)).

[15] The qualification regarding 'understanding a language' will not be further discussed in this paper, since no *externalist* theorist should want to deny that this criterion is satisfied. Indeed, it is one of the *virtues* of externalism that one can count as possessing a concept, and so as possessing (and in this sense *knowing*) a language whose lexicon expresses that concept, even when one's explicational knowledge regarding that concept varies over time.

[16] I thank Ray Elugardo for pointing out to me the need for condition (iii).

[17] Throughout the remainder of this paper I will use the expressions 'say the same thing' and 'express the same thought' interchangeably.

[18] Note that my conflation, of (a) intending to *say* the same thing and (b) intending to *mean* the same thing, does not involve conflating the concept *what S says* with the concept *what S means*. One can intend to mean something and yet say something other than what one intends to mean. Here what I am conflating is at the level of *intentions*; at this level, such a conflation is innocuous.

[19] This example, and much of the wording, were provided to me by Ray Elugardo.

[20] See Gareth Evans, "The Causal Theory of Names," *Aristotelean Society* 47 (1973).

[21] I should add that I do not endorse this reasoning. But for the sake of argument I am accepting the point that externalist considerations alone can undermine a thinker's intention to 'say the same thing' as she said on some previous occasion. I accept this in order to convince externalists who hold this view that STE allows the point that they would insist on.

[22] I thank Ray Elugardo for suggesting that I put the point in this manner.

[23] The phrase 'Judy does not know what she was thinking when she produced *W*' is meant to be a stylistic variation on 'Judy does not know the content of the thoughts she expressed with *W*.' The locution 'know *what*' one is thinking, as a way to spell out what 'knowledge of content' comes to, was suggested to me by Sidney Morgenbesser. I should point out, though, that Burge too uses this locution (Burge, 1988, p. 662), but in a way that is very different from mine; for Burge one counts as 'knowing what thought-content one is thinking,' in the sense relevant to knowledge of content, merely in virtue of satisfying the antecedent of SAKC.

[24] We might be reminded as well that this is the very idea at the heart of Falvey and Owens' (1994) conception of 'introspective knowledge of content proper,' as opposed to 'introspective knowledge of discriminatory content.' Their idea was that the former only involves true, justified self-ascription of a thought. In light of this, my argument against SAKC itself can be brought to bear against the way in which they would appeal to this distinction in contexts of the compatibilism debate.

[25] I defend this point at length, and draw out its implications, in "The Problem of Content-Identification" (unpublished manuscript).

[26] See Goldberg (1997).

[27] The one exception that comes to mind is Bilgrami (1992). (This is not an argument for incompatibilism *per se*, but only incompatibilism regarding authoritative knowledge of content with *certain 'orthodox' versions of* externalism.)

REFERENCES

Bilgrami (1992): 'Can Externalism be Reconciled with Self-Knowledge?', *Philosophical Topics* 20(1).

Boghossian (1989): 'Content and Self-Knowledge', *Philosophical Topics* 17(1).

Brueckner (1990): 'Skepticism of Knowledge of Content', *Mind* 99.

Brueckner (1997): 'Is Scepticism about Self-Knowledge Incoherent?', *Analysis* 57(4).

Burge (1988): 'Individualism and Self-Knowledge', *Journal of Philosophy* 85.

Burge (1989): 'Wherein is Language Social?', in George.

Evans (1973): 'The Causal Theory of Names', *Aristotelean Society* 47.

Falvey and Owens (1994): 'Externalism, Self-Knowledge, and Skepticism', *The Philosophical Review* 103(1).

George (1989): *Reflections on Chomsky*, Oxford: Basil Blackwell.

Gibbons (1996): 'Externalism and Knowledge of Content', *The Philosophical Review* 105(3).

Goldberg (1996): 'Introduction to Section IV, Entitled "Self-Knowledge"*;* in Pessin and Goldberg.

Goldberg (1997): 'Self-Ascription, Self-Knowledge, and the Memory Argument', *Analysis* 57 (3).

Goldberg (unpublished manuscript): 'The Problem of Content-Identification'.

Heil (1988): 'Privileged Access', *Mind* 97(386).

Jacobsen (1997): 'Self-Quotation and Self-Knowledge', *Synthese* 110.
Ludlow (1995a): 'Externalism, Self-Knowledge, and the Prevalence of Slow Switching', *Analysis* 55 (1).
Ludlow (1995b): 'Social Externalism, Self-Knowledge, and Memory', *Analysis* 55(3).
McKinsey (1990): 'Anti-Individualism and Privileged Access', *Analysis* 51(1).
Pessin and Goldberg (1996): *The Twin Earth Chronicles*, New York: M.E. Sharpe.
Warfield (1997): 'Externalism, Privileged Self-Knowledge, and the Irrelevance of Slow Switching', *Analysis* 57(4).

Department of Philosophy
Grinnell College
Grinnell, IA 50112
USA
(email: goldberg@ac.grin.edu)