

# *Self-ascription, self-knowledge, and the memory argument*

SANFORD C. GOLDBERG

## *1. Motivating the assumption: Burge on self-knowledge*

The thesis of this paper is that, in the context of an externalism about content, the assumption that

true justified self-ascription amounts to (or otherwise entails) self-knowledge of content

is tendentious. (Throughout this paper I shall refer to this claim as ‘the central assumption.’) Since I will be hanging much on a somewhat non-standard take on the self-knowledge debate, I will review the debate with an eye towards motivating the central assumption; in the section that follows I reconstruct an argument whose point (I claim) is to challenge that assumption.

Many people worry about the compatibility of the following two claims:

- (1) We have introspective and authoritative self-knowledge of the content of our intentional states (*the doctrine of authoritative self-knowledge*).
- (2) Some contents cannot be individuated in terms of properties of the individual considered in isolation from her social and physical environment (*the doctrine of externalism*).

This presents a problem for those who think that we have independent reasons to hold both doctrines.<sup>1</sup> Hence the reconciliation problem: to show that, despite appearances, these two doctrines can be reconciled.

Burge 1988 contains a proposal based on the observation that the same form of words which a person uses to *express*<sup>2</sup> a thought is also used to *self-ascribe* the thought. (I can express my thought that it’s raining by uttering ‘It’s raining,’ or I can self-ascribe this same thought by uttering ‘I think: it’s raining.’) His proposed reconciliation exploits the intimate connection between the expression and the self-ascription of thoughts in order to argue that there is no problem of authoritative self-knowledge for

<sup>1</sup> Theorists differ on what this problem is. For a review of three influential conceptions of the problem, see my introduction to the section entitled ‘Self-Knowledge’ in Pessin and Goldberg (1996).

<sup>2</sup> Nothing in Burge’s analysis hangs on the public expression of one’s thoughts; ‘expression’ can be read as neutral between thinking a verbalized thought to oneself and expressing it aloud.

the externalist. Burge points out (what is no doubt correct) that the doctrine of externalism does not affect the thinker's ability to express her thoughts verbally (by making the relevant utterances). Given this, he claims, we can be assured that there is no problem of authoritative self-knowledge for the externalist: the same conditions that individuate a thinker's first-order thought that *p* will enter into the individuation of her second-order thought that she is thinking that *p*. That this is so simply reflects the intimate connection between the expression and the self-ascription of thoughts.

Burge's gloss on these considerations is that self-ascriptions amount to *self-verifying judgements*. These judgements are self-verifying because one thinks the thought that *p* in the very act of thinking (or consciously reporting) that one thinks that *p*. Burge's central claim was that externalism does nothing to undermine this self-verifying nature of first-person reports. This manner of reconciliation provides us with the central motivation for equating self-knowledge with true justified<sup>3</sup> self-ascription. The motivation has two sources. The first is a felt need to provide the reconciliation. The second is the observation that a thinker is *always* in a position to self-ascribe a thought merely by uttering the words that express it. These two points make the central assumption attractive: insofar as it is acceptable, the assumption (that true justified self-ascription amounts to self-knowledge) would warrant Burge's conclusion that there is no problem of self-knowledge for the externalist.

## 2. *Challenging the central assumption: the memory argument*

I now want to review an argument – the Memory Argument (Boghossian 1989) – whose point is to challenge this very assumption. Though this argument has been criticized on grounds of relying on a questionable doctrine about memory, I suggest that the appeal to memory is eliminable. If the resulting argument is successful, it provides a reason to reject the central assumption.

<sup>3</sup> I have not yet said anything about the *justification* of these self-ascriptions. And, indeed, some critics of externalism have argued that even though an externalist can acknowledge the self-ascriptions as *true*, nonetheless there is a problem to show how the self-ascriptions can be *justified*. Such a charge is usually brought out by appeal to relevant alternative conceptions of justification in the context of Twin Earth examples. (Such an argument is examined in Boghossian 1989). I think that the charge that externalists cannot account for the justification of self-ascriptions, when brought out by appeal to relevant alternative considerations and Twin Earth examples, can be met. Indeed, Falvey and Owens 1994 present a rather compelling case for this. Consequently, I am simply going to *assume* for the sake of this argument that there is no problem of justification for the externalist. Since my ultimate aim is to suggest that nonetheless there is still a self-knowledge problem for the externalist, this assumption is concessive.

Some stage setting is necessary. Burge's account of self-knowledge was designed to show that, for any perceptual thought, *on the occasion of thinking that thought* the thinker knows its content in an immediate and direct way. Burge imagines an objection: suppose a person undergoes a series of slow switches between Earth and Twin Earth. Suppose further that the thinker was told of the switches at some *future* time, and then asked which thought she had at that earlier time. Burge's response involved (i) conceding that at that later time she 'may not know' which thought she had, while (ii) insisting that nonetheless she knew its content when she entertained it at that earlier time. The memory argument is meant to bring out the absurdity of the claim that 'although *S* will not know tomorrow what he is thinking right now, he does know right now what he is thinking right now' (Boghossian 1989: 22). To substantiate the untenability of such a position, Boghossian presents the memory argument:

... At any given moment in the present, say  $t_1$ , *S* is in a position to know what he is thinking at  $t_1$ . By Burge's criteria, therefore, he counts as having direct and authoritative knowledge at  $t_1$  of what he is thinking at that time. But it is quite clear that tomorrow he won't know what he thought at  $t_1$ . No self-verifying judgement concerning his thought at  $t_1$  will be available to him then. ... But there is a mystery here. For the following would appear to be a platitude about memory and knowledge: if *S* knows that *p* at  $t_1$ , and if (at some later time)  $t_2$ , *S* remembers everything *S* knew at  $t_1$ , then *S* knows that *p* at  $t_2$ . Now, let us ask: *why* does *S* not know today whether yesterday's thought was a *water* thought or a *twater* thought? The platitude insists that there are only two possible explanations: either *S* has forgotten or he *never* knew. But surely memory failure is not to the point. ... The only explanation, I venture to suggest, for why *S* will not know tomorrow what he is said to know today, is not that he has forgotten but that he never knew (1989: 22–23).

One feature of this argument is notable. A proponent of the memory argument need not (indeed, should not) deny that

*S* at  $t_1$  is in a position to form a justified self-ascription of the thought that he is thinking at  $t_1$ .

This is just a special case of the general point about the intimate connection between the *expression* and *self-ascription* of thoughts, which goes unchallenged by the argument. What the argument purports to show is not that *S* can fail to be able to form a justified self-ascription of a thought, but rather that there are cases in which

*S* does not know at  $t_1$  the thought he has at  $t_1$ .

Since this is (purportedly) established on grounds that leave the point about self-ascription intact, the memory argument is a perfect candidate for challenging the central assumption.

However, the argument as Boghossian formulates it has been criticized as relying on a controversial doctrine about memory. For example, Ludlow (1995) offers a helpful formulation of the argument:

- (1) If  $S$  forgets nothing, then what  $S$  knows at  $t_1$ ,  $S$  knows at  $t_2$ .
- (2)  $S$  forgot nothing.
- (3)  $S$  does not know that  $p$  at  $t_2$ .
- (4) Therefore,  $S$  does not know that  $p$  at  $t_1$ .

Ludlow then challenges (1) by suggesting that no externalist ought to accept this doctrine (he goes on to propose a new conception of memory, motivated by his prior acceptance of externalism). Nor is (1) objectionable merely by the lights of the externalist. Falvey (in conversation) has suggested that (1) is objectionable as a general principle about memory, on the (largely uncontroversial) grounds that motivate the point that is at the heart of Goldman 1976: someone could know that  $p$  at  $t_1$ , remember at  $t_2$  everything she knew at  $t_1$ , and yet fail to know that  $p$  at  $t_2$  – *even if* she continues to believe that  $p$ , and  $p$  is true – for the very familiar reason that there might be *new evidence* encountered along the way that points to a relevant alternative she cannot exclude. If this is right, then (1) is not acceptable.

I propose to grant that the memory principle will not do, and to respond by revising the Memory Argument so that it no longer relies on such a principle. To do so I will de-emphasize the *memory* aspect of the Memory Argument, and will underscore the *point* that the Memory Argument used judgements involving memory to bring out. If the Memory Argument is not essentially about memory at all, then we should be able to reach Boghossian's conclusion in (4) without relying on the controversial premise in (1); and indeed I try to show just this.

In order to show that the memory argument is misnamed – that it is not about memory in the first instance at all, and that its conclusion does not depend on an appeal to the objectionable memory principle – we can change Boghossian's story a bit. Let  $S$  have a (verbalized) thought in her mind all the while, and suppose that as she is thinking it she overhears someone telling the story of her ( $S$ 's) world-switching. Moreover, suppose that she does not know (and is not told) whether she is presently on Earth or Twin-Earth. Presumably, this new way of telling the story neither introduces new variables nor affects any of those in the original story. (After all, the postulated difference in time between  $t_1$  and  $t_2$  is merely a stylistic feature of the original story, meant to make vivid the shift in epistemic

status of *S*'s second-order thought *before* and *after* she learns of her world-switching history.) Now, Burge was willing to concede the point about a possible failure of knowledge of her thought in the case where  $t_2$  (the time of 'recollection') is later than  $t_1$  (the time of the original thinking of the thought); since the new example is just a stylistic variant on the same story, parity of reasoning requires him to concede the same point (about a possible failure of knowledge) where there is no difference in time at all. But then he would be conceding that upon hearing this story *she does not know the thought she has*, even as she thinks it to herself!

To vindicate my claim that the argument is at bottom not about memory after all, I propose replacing the contested memory principle in (1) with the following *Principle of Knowing Identification* (PKI):

- (1') 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 no presently available way to select 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 thought in question.

This transforms the argument from a point about memory to a point about the difference between being able to *self-ascribe* a thought and being able to *identify knowingly* the thought self-ascribed. The claim would then be that for the memory case Boghossian describes

- (2') *S*'s self-ascription of the thought that *p* satisfies (i) and (ii).

The conclusion is the same: *S* does not know the thought she has at  $t_1$  (even though she is able to self-ascribe the thought at  $t_1$ ).

Though PKI is rather different from Boghossian's memory principle, I believe that it captures the point Boghossian's appeal to memory was meant to bring out. Without arguing for this, however, I want to give examples that are meant to show the plausibility of the principle.

The clearest example would be the case of a thought involving a word that is genuinely ambiguous, where one has forgotten which meaning she had in mind. Suppose *S* recalls having expressed an earlier thought with the utterance 'The bank is about four blocks from the station,' yet does not remember if she had an effluvial embankment or a financial institution in mind. In such a case it would be perfectly unintuitive to say that *S* could knowingly identify that thought merely by using the same form of words she'd used earlier, even if *S* were to use these words with the intention that they mean whatever they meant as used to express the original thought.

Nor is this a restricted example. The same point can be made by considering thoughts in whose expression a proper name occurs. Suppose *S* remembers having had a thought she expressed by ‘Mary is in town,’ but fails to remember which of the several Marys she knows she was thinking about. Again, it would be most unintuitive to say that *S* can knowingly identify her thought merely by uttering the same form of words with the suitable (indexically-specified) intention.

While both of these examples are memory-involving ones, the point they support is not essentially about memory at all. The claim is that Twin-Earth cases exhibit the same relevant features as these cases, but without the memory component. In each, there is a gap between the use of words to self-ascribe a thought (even as supplemented by the right indexically-specified intention), and the knowing identification of the thought thereby expressed. It is the point of the revised Memory Argument that one can be in a position to self-ascribe without knowing the thought thereby ascribed, *even as one does so*.

One final comment about what these examples show will serve to indicate how strong the reformulated argument actually is. Recall what Burge was willing to concede in the original case: that

*S* may not know at  $t_2$  the thought she had at  $t_1$  merely in virtue of having found out about her world-switching.

In conceding this, Burge resisted what (from the perspective of one trying to secure externalism against charges of jeopardizing self-knowledge) must be a tempting move. In particular, he resisted the temptation to insist on a doctrine that I shall call the ‘strong identification,’ the thesis that

*S* can identify at  $t_2$  the thought she had at  $t_1$ , simply by using the *same form of words* as she used when she self-ascribed the thought at  $t_1$ .<sup>4</sup>

Had he insisted on this, he would not have been conceding the point that the Memory Argument (in either of its incarnations) requires to get going. In light of this, it appears that he ought to have insisted on the strong identification. However, if I am correct to assimilate the Twin Earth ‘memory’ case into the same category as the ambiguity or proper name cases just discussed, then it is clear that Burge was correct to resist the strong identification (for the reasons given above).

In fact, Burge’s decision not to insist on the strong identification can be read as an implicit acknowledgement of the power of PKI. On my view, he is implicitly acknowledging the plausibility of the claim that, once informed of her world-switching, *S does not know at  $t_2$  the content of (some of) the words she has been using* – at least as they were used in the

<sup>4</sup> Again, we might supplement this with the proviso: ‘so long as *S* intends that the words mean whatever they meant for her at  $t_1$ ’.

previous self-ascription of the thought. That is, once *S* learns of the world-switching, the term ‘water’ comes to be ambiguous for her<sup>5</sup> between *water* and *twater*, in such a way that a re-mouthing of the same form of words would not suffice to identify knowingly the thought that she had.

Now, it might be thought that PKI amounts to a contentious conception of self-knowledge of content, one that begs crucial questions against Burge. In particular, Burge and his followers have been most careful to distinguish between the doctrine of self-knowledge of one’s thought from other doctrines with which the former is often confused but which (by their lights) are nonetheless distinct. Chief among these are conceptions of self-knowledge on which

To know one’s own thought is to be able to give a complete explication of the concepts that figure in it.<sup>6</sup>

and

To know one’s own thought is to be able to discriminate it from any other content-distinct thought.<sup>7</sup>

Is PKI guilty of smuggling one of these doctrines into its conception of self-knowledge? In the space that remains I want to argue that, though PKI does require some discrimination between thoughts with different contents, nonetheless the sort of discrimination it insists on (i) does not amount to a requirement of complete explication and (ii) is independently motivated. I will bring these out in reverse order.

Return to the example involving the ambiguity of ‘bank.’ I suggested that, assuming that *S* does not remember which of the disambiguations she had in mind, it is most natural to deny that *S* knowingly identifies her thought merely by using the same form of words with the suitable indexically-specified intention. What accounts for this? Well, she would not be credited with knowing her thought *unless she had some way for determining which of the two thoughts* – a financial-institution-thought or an side-of-river-thought – she had; and for this a re-mouthing of words would not be enough. In

<sup>5</sup> Actually, stating the ambiguity here requires some care. It is possible that even after being told of her world-switching all *S* knows is that ‘water’ as she used it previously was used to denote two liquids with very different chemical substructures. If she does not know any chemistry at all, then it may not make sense to call this an ambiguity between *H<sub>2</sub>O* and *XYZ* (italics used to denote the *meanings* of these expressions, not their referent).

<sup>6</sup> In his 1987 p. 662 Burge writes, ‘One should not assimilate ‘knowing what one’s thoughts are’ in the sense of basic self-knowledge to ‘knowing what one’s thoughts are’ in the sense of being able to explicate them correctly.’ See also Burge 1989, where he sharply distinguishes between a ‘lexical item’ and ‘the explication of its meaning.’

<sup>7</sup> This distinction, between introspective knowledge of content and introspective knowledge of comparative content, is exploited in Falvey and Owens 1994.

other words, in the ambiguity case, the view that knowledge of thought requires a minimal sort of discrimination is not contentious in the least.

Nor does this insistence on discrimination amount to a complete explication requirement. *S* need not be able to explicate exhaustively the concept of *bank* in order to be able to identify knowingly the thought she had earlier; she would merely have to know enough to make clear that she meant financial institution (not side of a river). For this it would suffice for her to say something like ‘the place where one takes out money’. Whether *S* could then go on to provide a comprehensive explication of the concept *bank* (as distinct from concepts for other places where one takes out money) is a substantial, though from the present perspective irrelevant, point. This makes clear that one can satisfy PKI without having to satisfy the complete explication requirement.

In fact, the plausibility of PKI can be brought out by a look at the *overly-liberal* conception of self-knowledge that PKI is meant to replace. According to this conception,

a sufficient condition for *S* to identify knowingly an earlier thought is that (1) she remember the form of words she used to express it, and (2) she is in a position to form the intention to use the words so as to mean whatever they meant for her on the past occasion in question.<sup>8</sup>

That this conception is overly-liberal is seen quite clearly in the ambiguity and proper name cases. I have been urging that this same point – the over-liberality of this conception – holds in the world-switching scenario. Consequently, if one is unwilling to allow that a re-mouthing of words backed by an indexically-specified intention suffices as a knowing identification of the thought expressed in the ambiguity and proper-name cases, then one ought to be similarly unwilling to allow this in the case Boghossian discusses. PKI is meant to be the principle that indicates what these cases have in common.

I want to conclude by way of saying what the present argument owes to the original formulation of the Memory Argument. The revised argument follows the strategy of its ancestor in trying to make intuitive the notion that (given externalism) there is more to knowledge of one’s own thought than correct justified self-ascription. Here the significance of the appeal to memory *is* worth noting, for what it makes vivid. That one can raise the question *Does S know her thought?* in the context of *past* thoughts<sup>9</sup> reveals that an exclusive focus on self-ascriptions of *occurrent* thought distorts

<sup>8</sup> Obviously, the indexical specification of the intention is central.

<sup>9</sup> I think there is at least one other context as well in the context of assessing the inferences drawn (or not drawn) in practical syllogisms. On this see Loar 1985 and Bilgrami 1992.



one's understanding, both of the self-ascriptions themselves as well of the nature of self-knowledge claims. The central assumption is symptomatic of this artificially narrow focus; or so a proponent of the revised argument might suggest.

### 3. Conclusion

In this paper I have suggested that Burge's externalist account of self-knowledge depends on the assumption that true justified self-ascription of content amounts to (or entails) self-knowledge of content, and that this assumption is challenged by a proper reformulation of Boghossian's Memory Argument. My conclusion is conservative: in the context of a question about the compatibility of externalism and authoritative self-knowledge, this assumption is tendentious. This is meant as a challenge to those interested in securing the mentioned compatibility: if such is to be had, the externalist must do so by a principled rejection of the line of argument sketched above. The 'memory' argument as reconstructed here does not appeal to any contentious conception of self-knowledge; instead it turns on a plausible principle about the difference between self-knowledge and the re-mouthing of words, as backed by a rejection of what is in any case an overly liberal conception of that knowledge.<sup>10</sup>

Box V-5, Grinnell College  
Grinnell, IA 50112, USA  
goldberg@ac.grin.edu

### References

- Bilgrami, A. 1992. Can externalism be reconciled with self knowledge? *Philosophical Topics* 20: 233–67.
- Boghossian, P. 1989. Content and self-knowledge. *Philosophical Topics* 17: 5–25.
- Burge, T. 1988. Individualism and self-knowledge. *Journal of Philosophy* 85: 649–63.
- Burge, T. 1989. Wherein is language social. In *Reflections on Chomsky*, ed. A. George, 175–91. Oxford: Basil Blackwell.
- Falvey, K. and J. Owens. 1994. Externalism, self-knowledge, and skepticism. *The Philosophical Review* 103: 107–137.
- Goldman, A. 1976. Discrimination and perceptual knowledge. *The Journal of Philosophy* 76: 771–91.
- Loar, B. 1985. Social content and psychological content. In *Contents of Thought*, ed. R. Grimm and D. Merrill, 99–110. Tucson: University of Arizona Press.
- Ludlow, P. 1995. Social externalism, self-knowledge, and memory. *Analysis* 55: 157–59.
- Pessin, A. and Goldberg, S. 1996. *The Twin Earth Chronicles*. New York: M. E. Sharpe.

<sup>10</sup> I would like to thank Akeel Bilgrami, Kevin Falvey, Sidney Morgenbesser, and Stephen Schiffer for their helpful comments on earlier drafts; and Adam Vinueza for a helpful discussion of these matters.