EXTERNALISM, INCLUSION, AND KNOWLEDGE OF CONTENT*

Carlos J. Moya. University of Valencia

In this paper I shall address the question whether self-knowledge is compatible with an externalist individuation of mental content. This question becomes pressing only in so far as self-knowledge is taken to be a genuine cognitive achievement. Not everybody accepts this. Some conceptions of self-knowledge interpret first-person statements about mental states as expressive, non-cognitive statements. Wittgensteinian approaches take this line. On this perspective, the privilege of the first-person amounts to an entitlement to express mental states: linguistic statements replace more primitive, non-linguistic expressions. No cognitive achievement is involved here. Related positions view first-person avowability as constitutive of the human mind itself. Against these approaches, I take self-knowledge to be a genuinely cognitive achievement: it is correct, I think, to say that we usually know what we currently want, believe or intend. Against a Cartesian perspective, self-knowledge is neither incorrigible nor infallible. It is, however, direct, in the sense of non inferential, a priori, in the sense of not being based on empirical investigation of one's surroundings, and, in normal cases, presumptively true and endowed with prima facie special authority. Now the question is: Can we have self-knowledge, so understood, if we also accept externalism, i.e., if we accept that mental content is

* Research for this paper has been funded by the Spanish Government's DGES as part of the project PB96-1091-C03-02. My thanks to this institution for its help and encouragement. I would also like to express my gratitude to Sven Bernecker, Tobies Grimaltos, Andreas Kemmerling and Nenad Miscevic for their help and comments on previous drafts of this paper.

constitutively determined by factors distinct from and external to the subject's brain or body? My answer to this question will be affirmative. There are, as far as I know, two main lines of attack to compatibility between self-knowledge and externalism. According to one of them, compatibilism entails that a subject can know, a priori, substantial truths about the external world.[1] The other line rests on the possibility of a subject's being unwittingly switched between worlds (cf. Boghossian 1989, 1992, 1994). In this paper I shall consider only the latter, though I think that my proposal could also be shown to be successful against the former.[2] The philosophical interest of the question addressed here lies in the fact that both externalism and self-knowledge seem to be true and important doctrines about the human mind. Without externalism, we cannot make distinctions between thought contents (e.g., between water-thoughts and twater-thoughts) which are intuitively there, nor can we put forward certain intuitively correct semantic evaluations (e.g., that water-thoughts are false on Twin Earth). Besides, externalism contains the promise of a picture of the mind that could set us free from the traps of Cartesianism. Self-knowledge, in turn, is still more central to our conception of mentality. Lack of self-knowledge threatens responsible agency and critical rationality, as some authors have rightly stressed (cf. e.g. Bilgrami 1992, pp. 250-1 and Burge 1996). So, it would certainly be good news if we could have both externalism and self-knowledge. Let us see whether we can.

---

[1] The original argument was first put forward in McKinsey 1991. See also Brown 1995. The argument is given a new version in Boghossian 1997. For replies to Boghossian's version see my 1998 and Brown 1999. For recent discussions of this line of argument see Brueckner 2000 and Falvey 2000.

[2] As I try to argue in my 1998 paper.

1. The Simple Argument.

At first sight, there is at least a tension between the thesis that content is determined by external factors on which a subject has no special cognitive authority and the thesis that a subject has a special cognitive authority over this content. This tension can be given expression in what I shall call "the simple argument". The argument can be put as follows: if the contents of one's thoughts depend on factors over which one has no direct, a priori, authoritative knowledge, one cannot have direct, a priori, authoritative knowledge over those contents themselves (first premise); but the antecedent of this conditional is true if externalism is true (second premise); so, if externalism is true, one cannot have self-knowledge (conclusion). This argument is implicit in many incompatibilist authors (cf. Woodfield 1982, pp. vii-viii). But we can find it explicitly stated by Laurence Bonjour in a reference book, the *Blackwell Companion to Epistemology*. He writes: "An objection to externalist accounts of content is that they seem unable to do justice to our ability to know the contents of our beliefs or thoughts 'from the inside', simply by reflection. If content is dependent on external factors pertaining to the environment, then knowledge of content should depend on knowledge of those factors —which will not in general be available to the person whose belief or thought is in question." (Bonjour 1992, p. 136). That this text appears in this important reference book is a symptom that incompatibilism is gradually becoming the received opinion on this subject and that the simple argument for it seems convincing to many thinkers.

And, nevertheless, the simple argument does not prove the incompatibility thesis, for, even if the argument is valid, its first premise is arguably false. To see this, think that there is a dependence

relation, maybe of a metaphysical character, between my existence and my parents' existence, so that I would not exist if my parents had not existed. However, I know that I exist in a direct, authoritative, a priori way, whereas my knowledge of my parents' existence is only empirical (cf. McKinsey 1991; see also Heil 1992, p. 163). By the way, this is why Descartes' argument from the certainty of the Cogito to the conclusion that the thinking self is independent of the body fails, as Arnauld already noticed (cf. Burge 1988, p. 651). It seems, then, that the first premise of the Simple Argument is false. At least, it cannot be true as a case of the general statement according to which if A depends upon B, one cannot have direct, a priori, authoritative knowledge of A if one does not have such a knowledge of B. The case of my parents' and my own existence is a counterexample to this statement. So, content might also be known in a direct, a priori, authoritative way even if it depends on external factors that cannot be known that way. If A depends upon B, it is not true, in general, that, in order to know A, I have first to know B. This holds as well when the dependence relation is constitutive or conceptual. There is such a relation between a certain figure's being a triangle and its internal angles' measuring 180 degrees. However, I can know in a direct, authoritative way that a certain figure is a triangle without knowing that its internal angles measure 180 degrees. We can conclude, then, that the simple argument does not prove that self-knowledge and externalism are incompatible. Incompatibilists must work harder in order to substantiate their claim.

2. The Inclusion Model.

The failure of the Simple Argument does not dispel the feeling that there is a conflict or tension between self-knowledge and externalism. A positive account of how they are compatible would be welcome. The most widely accepted compatibilist account of self-knowledge is what Sven Bernecker has called "the inclusion theory of self-knowledge" (Bernecker 1996, p. 265) and I would prefer to call "the inclusion model" of self-knowledge. Advocates of this model include Tyler Burge, Donald Davidson and John Heil, among others.[3] As Sanford C. Goldberg has pointed out, this proposal exploits the fact that the same form of words used to express a thought (e. g. "it's raining") is also used to self-ascribe the thought (e. g. "I think: it's raining") (Goldberg 1997, pp. 211-212). The central idea of this model is that reflective awareness of a first-order thought (say, a reflective judgment that one has a certain thought) inherits or includes (whence the label "inclusion model") the content of the first-order thought itself. Burge writes: "... Knowledge [of one's own mental events] consists in a reflexive judgment which involves thinking a first-order thought that the judgment itself is about. The reflexive judgment simply inherits the content of the first-order thought." (Burge 1988, p. 656; cf. also Heil 1988, p. 246 and 1992, ch. 5). On an externalist perspective, the individuation conditions of certain thoughts are partly external to the thinker; they consist in certain links with some aspects of the environment or with a social linguistic community to which the thinker defers. These conditions enable me to have a thought with a certain content, even if I do not know or believe that these conditions obtain or even what they are. Now, when I reflexively ascribe that thought to myself, the content of this thought is simply included in my reflexive

[3] For a recent defence of a refined version of this model see Gibbons 1996.

awareness, and the external conditions that partly determine the content of the first-order thought also contribute to determining the content of the reflexive or second-order thought. As Davidson points out: "Showing that there is no conflict [between externalism and knowledge of content] is basically simple. It depends on realizing that whatever is responsible for the contents of our thoughts, whether known or not, is also responsible for the content of the thought that we have the thought" (Davidson, unpubl. ms, p. 35). The inclusion model ensures that Cogito-like judgments are contextually self-verifying. Suppose, in effect, that external conditions partly determine, for the thought I express with "water quenches thirst", the content that water quenches thirst. I may not know that those conditions do in fact obtain or, for that matter, what they are, but if I am reflexively aware of this thought and I express this reflexive awareness with, say, "I am hereby judging that water quenches thirst", the content of the that-clause of this reflexive attitude is that water quenches thirst, that is, it is the content of the first-order thought itself. But if a corresponding episode happens on Twin-Earth, the thought that my Twin expresses with "water quenches thirst" has the content that twin water (twater) quenches thirst, and this is also the content of the that-clause of my Twin's reflexive attitude. So, on the inclusion model, Cogito-like judgments are reliably true in that they are contextually self-verifying. In this sense, a subject can be said to have authoritative knowledge over his thoughts' content, in spite of these thoughts' having external individuation conditions which the subject has no privileged knowledge of.

I think that the inclusion model is ultimately correct, but, as it stands, it is affected by some shortcomings. One problem is that it draws too heavily on an externalist, reliabilist view of justification (see,

e. g., Bernecker 1998), for, on this model, justification of self-ascriptions rests on the existence of a mechanism (which the subject need not have cognitive access to) ensuring that the content of first-order thoughts is ipso facto included in self-ascriptions of those thoughts. This important reliabilist component links the inclusion model too closely to the fate of reliabilism itself. It also needs some more work of a general kind about the nature of external determination of meaning and content if it is to be able to meet the objections we are about to see. One major objection is that this model does not ensure what Boghossian has called the "transparency of content". I shall try to do part of the required additional work later in this paper. Let us now turn to Boghossian's objection.

3. Transparency of content: switching between worlds.

Some philosophers feel that the inclusion model makes self-knowledge into a rather anaemic cognitive achievement, so much so that its entitlement to the dignity of knowledge could be justifiably questioned. Paul Boghossian, for one, acknowledges that accounts of self-knowledge on the lines of the inclusion model ensure that, at least in basic cases, a subject's reports of his current thought contents will always be true (Boghossian 1992, p. 15), but he complains that such accounts do not ensure that these thought contents are transparent for him. Boghossian construes the notion of transparency of content on the basis of Dummett's concept of transparency of meaning. According to Dummett, "... meaning is *transparent* in the sense that, if someone attaches a meaning to each of two words, he must know whether these meanings are the same" (quoted from Boghossian 1992, p. 16). Correspondingly, content is transparent to a subject only if he is able to

know in a direct, a priori way whether the contents of two thoughts of his are the same or not. Boghossian grounds his contention that the inclusion model does not grant transparency of content on thought experiments in which we are asked to imagine that a subject is unwittingly transported, say, from Earth to Twin Earth and remains there for a fairly long time.[4] Let's call our inter-world traveller "Peter". Boghossian writes: "How should we think about the semantics of Peter's thoughts? Well, one intuition that is shared by practically everyone who has thought about these cases is that, after a while (how long is unclear), tokens of 'water' in Peter's mentalese will cease to mean *water* and will come to mean *twater*" (Boghossian 1992, p. 18). This intuition coheres especially well, according to Boghossian, with the following principle of content fixation which underlies standard Twin Earth cases: "The contents of thought tokens of a given syntactic type are determined by whatever environmental property is the typical cause of the perceptions that cause and sustain tokens of that type" (Boghossian 1992, p. 19). I shall dispute later both the intuition and the principle, but let us provisionally grant them for the sake of the argument. Consequences of this thought experiment for self-knowledge are quite clear. Suppose, in effect, that, while still on Earth, in summer, Peter is thirsty and fills a glass with cold water while muttering: "Water will quench my thirst". Some years later, a similar episode takes place on Twin Earth and Peter mutters tokens of the same words, while remembering the analogous occasion we have referred to. Since his subjective experience has not been disrupted and, for all he knows, he has not travelled to another world, he will certainly judge that the

[4] This sort of thought experiment had been already devised by Burge: see Burge 1988, p. 652.

token thought contents he has expressed on both occasions are of the same type. But, if externalism is true, they are not. So, content is not transparent for him. His comparative judgment about his thought contents is false. On this basis, Boghossian contends that Burge's self-verifying judgments do not amount to knowledge. He tries to show this as follows. Suppose that, after being on Twin Earth long enough, Peter is told that the switch has occurred, but not when it took place. We ask Peter: "Two years ago, were you thinking that water quenches thirst or that twater quenches thirst?" Peter will not know the answer. However, according to Burge's inclusion model, two years ago Peter knew what he was thinking in that he was able to reflexively self-ascribe a thought he expressed with the sentence "water quenches thirst". But now Peter acknowledges that he doesn't know what thought he was having two years ago. Why? According to Boghossian, there are two possible explanations: that Peter has forgotten or that he never knew. But memory failure should be excluded by stipulation, for "it is not as if thoughts with widely individuated contents might be easily known but difficult to remember. The only explanation, I venture to suggest, is not that he has forgotten but that he never knew. Burge's self-verifying judgments do not constitute genuine knowledge. What other reason is there for why our slowly transported thinker will not know tomorrow what he is said to know directly and authoritatively today?" (Boghossian 1989, p. 23). Therefore, externalism is not compatible with self-knowledge. This is, in rough terms, Boghossian's incompatibilist argument.

4. Does self-knowledge include transparency of content?

One possible way of countering Boghossian's incompatibilist argument is to hold that, appearances to the contrary notwithstanding, ordinary self-knowledge does not include transparency of content. This thesis has been defended by Kevin Falvey and Joseph Owens.[5] Falvey and Owens contend that, independently of externalism, we do not enjoy what they call "introspective knowledge of comparative content". I take this knowledge to be equivalent to Boghossian's transparency of content, for it is characterized as follows: "With respect to any two of his thoughts or beliefs, an individual can know authoritatively and directly (that is, without relying on inferences from his observed environment) whether or not they have the same content" (Falvey and Owens 1994, pp. 109-110; cf. also Owens 1995). They seem to endorse the inclusion model of self-knowledge, with its resulting self-verifying judgments, and, on this basis, they accept that we possess what they call "introspective knowledge of content", according to which "an individual knows the contents of his occurrent thoughts and beliefs authoritatively and directly (that is, without relying on inferences from observation of his environment)" (Falvey and Owens 1994, pp. 109-110).

Boghossian's objection is quite probably innocuous against introspective knowledge of content. So, an externalist might easily embrace compatibilism by limiting the scope of self-knowledge in this way. But I have the suspicion that this would be too cheap a victory against incompatibilism. For, on the one hand, Falvey and Owens defend his thesis that we do not enjoy comparative self-knowledge on the basis of examples involving pairs of synonymous and co-extensive terms (e.

---

[5] This view has also been defended by John Gibbons. See Gibbons 1996, p. 304. It is also present in Burge 1998.

g. 'physician' and 'doctor', or 'cilantro' and 'coriander') which a subject does not know to be so. But these examples seem to me controversial and capable of receiving quite natural interpretations which do not entail that a subject lacks comparative self-knowledge. Suppose, for instance, that Andrew thinks that the thought he expresses with the words "Mary is a physician" is not the same as the thought he expresses with "Mary is a doctor". Suppose he thinks so because he believes that 'physician' and 'doctor' are not synonyms (let us suppose that he thinks that 'physician' is synonymous with 'physicist'). Only from an implausibly crude externalist perspective would someone be inclined to think that the subject is wrong about his two thoughts' being different. I tend to hold that Andrew's two thoughts are in fact different, that he is right in believing that they are and, therefore, that he has comparative self-knowledge. And, on the other hand, Boghossian's transparency requirement for self-knowledge seems to be well grounded on a reasonable requirement (a necessary condition) for knowledge in general. This requirement might be called "the principle of relevant alternatives". According to this principle, a subject cannot be said to know that *a* is an *F*, though it in fact is, if, in case *a* were a *G*, where being a *G* is a relevant alternative to being an *F*, this subject would still judge that *a* is an *F*. Knowledge requires the ability to discriminate between relevant alternatives. So, imagine, to use the famous Putnamian example, that I cannot tell elms from beeches. Then, even if my judgments of the form "*a* is an elm" happen to be always true, I cannot be said to *know* that *a* is an elm, for, if *a* were a beech instead, where being a beech is a relevant alternative to being an elm, I would still judge that it was an elm. Note that this still holds even if I do not know about the existence of beeches. It is the existence of relevant

alternatives, not my belief or knowledge that there are, what matters here for possession of knowledge. Now Peter, the inter-world traveller, does not seem to satisfy this requirement. He cannot discriminate between his water-thoughts and his twater-thoughts and the latter are, for him, relevant alternatives to the former. He does not know about the existence of these relevant alternatives, but, as I said, this does not change matters. In terms of the inclusion model, Peter's judgments about his thought contents happen to be true. When Peter is on Earth and reflexively mutters "I am judging that water quenches thirst", this judgment is true, for it has the content that water quenches thirst, but Peter does not *know* that his thought has this content, for, if it had the content that twater quenches thirst instead, he still would think, mistakenly, that it had the same content it now has.

This, if correct, restates Boghossian's incompatibilist argument against Falvey and Owens' response and makes things harder for compatibilists. They had better show that externalism is compatible with transparency of content if they want to show that it is compatible with self-knowledge.

5. Incompatibilism and memory.

Other attempts to meet Boghossian's objection focus on the role memory plays in his argument. Let us extend a bit on this part of the argument. Recall that Peter, the inter-world traveller, interpreted in the light of the inclusion model and its corresponding self-verifying judgments, knows at t1 what he is thinking, but at t2, after being told about the switch, he does not know what he was thinking at t1. But Boghossian takes it to be a "platitude about memory and knowledge" that "if S knows that p at t1, and if at (some later time) t2, S

remembers everything he knew at t1, then S knows that p at t2" (Boghossian 1989, p. 23). Now if S does *not* know that p at t2, then either he does not remember at t2 everything he knew at t1 or he did not know that p at t1. If, by assumption, we rule out memory failure, it seems we have to conclude that S did not know that p at t1. Ludlow has given this useful reconstruction of this part of Boghossian's incompatibilist argument: "(1) If S forgets nothing, then what S knows at t1, S knows at t2, (2) S forgot nothing, (3) S does not know that P at t2; (4) therefore, S did not know that P at t1" (Ludlow 1995a, p. 157).

That the 'platitude' Boghossian states (which corresponds to premise (1) in Ludlow's reconstruction) is indeed a platitude has been put into question by some authors, such as Ludlow (1995b) and Goldberg (1997). Goldberg's remarks are especially interesting from the perspective of the present paper, for they point to an element that seems to be an integral part of Boghossian's argument, namely the relevant alternatives account of knoweldge, in order to undermine Boghossian's supposed platitude. According to Goldberg (who acknowledges his debt to Falvey on this point), "someone could know that *p* at *t1*, remember at *t2* everything she knew at *t1*, and yet fail to know that *p* at *t2* —*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" (Goldberg 1997, p. 214). Brueckner (1997) has also questioned premise (1) on a similar basis. He writes: "To say that at t2, S has forgotten nothing that he knew at t1 is to say that at t2, remembers everything that he knew at t1. But it does not follow that that at t2, S knows everything that he knew at t1. This is because for some P that he knew at t1, he may remember at t2 that P while failing to

know at t2 that P, say, because of some defeating information that he has learned between t1 and t2" (Brueckner 1997, p. 8). In Boghossian's example, Peter has acquired new information (namely that he has been switched at some unknown time) which defeats his justification for believing that at t1 he was thinking that water quenches thirst. This, if correct, shows premise (1) to be false.

Is Boghossian's incompatibilist argument thereby defeated? I do not think so, for the argument can be restated without any essential appeal to memory. The conclusion that Peter does not know at t1 what he is thinking at t1, so that Burge's self-verifying judgments do not constitute knowledge, can be reached with no need of premise (1). Suppose, in effect, that Peter is told that he has suffered repeated switching between Earth and Twin Earth, but not when the switches took place nor where he is now. In these circumstances, he will recognize that he does not know, right now, what he is thinking right now, because he does not know whether he is thinking that water quenches thirst or that twater quenches thirst.

The possibility that Boghossian's argument is restated without appealing to memory suggests that no criticism based on reflections on memory can aspire to a definitive rebuttal of Boghossian's incompatibilism. Goldberg has also tried to show that Boghossian's argument may dispose of an appeal to memory and be restated in terms of an ability to knowingly identify self-ascribed thoughts (Goldberg 1997, p. 215). Both Goldberg's and my own restatement of Boghossian's incompatibilist contentions without appeal to memory make use of the notion of discrimination between thought contents, a notion closely related to that of transparency of content and to the relevant alternatives account of knowledge. I have argued in favour of

transparency of content as a necessary condition for self-knowledge in the previous section, against Falvey and Owens' denial of this condition. I want now to consider a possible objection to this condition, according to which it might seem implausibly strong to require of Peter, as far as he is fully unaware of his switching career, that he be able to knowingly identify his *water*-thoughts as *water*-thoughts and his *twater*-thoughts as *twater*-thoughts. This form of discrimination we might call "strong discrimination". I can agree that this requirement is too demanding. However, it is not implausible to require of Peter, if he is to have self-knowledge, at least that he does not take his *water*-thoughts and his *twater*-thoughts to be of the same type. We might call this "weak discrimination". Now, Peter does not satisfy even the weak discrimination condition. If he is unaware of the switching, he will certainly judge that a *water*-thought he expresses with 'water quenches thirst' is of the same type as a *twater*-thought he expresses with those same words, even if he is having those two thoughts in the specious present (provided both concepts are available to him). The inclusion model ensures that his self-ascriptions are true, but not, if the weak discrimination condition is correct, that they amount to knowledge.

In a recent paper (Burge 1998),[6] Burge has countered Boghossian's argument by distinguishing two ways in which memory can work, namely by discriminatory identification and by preservation. Preservative memory retains the content of past thinkings through causal links with them, preserving contents and attitudes without the subject's having to refer to or identify them, similarly to what happens with anaphoric uses of pronouns. So, by simply having a thought that

[6] Some central ideas in this paper relate closely to those Burge had developed in his 1993 paper.

Peter expresses with "I thought that water is wet", preservative memory, working properly, connects Peter's present thought to the remembered though, with the content and concepts that were in play at that earlier time.

Burge holds, then, that discriminatory identification is not necessary in order to know what one is thinking. Besides, a strong reliabilist component is present in Burge's account of presevative memory, as is also in his understanding of present-tense self-knowledge. These two aspects may undermine the force of his response to Boghossian, if either reliabilism were shown to be wrong or if the ability to discriminate between relevant alternatives were to be a necessary condition for knowledge. I have argued in favour of the latter above. Suppose that Peter, being very thirsty on Twin Earth, thinks to himself "water will quench my thirst" and self-ascribes this thought, while also thinking "I thought the same thing two years ago". It is hard to accept that in these conditions, where the comparative thought is supposedly false, his plain self-ascription amounts to knowledge. But then it is hard (and arbitrary) to accept that, in order to have self-knowledge, the self-ascribed thought must not be accompanied by a corresponding comparative thought.

It would be good, then, to have compatibilism even accepting the discrimination condition. And it would also be good if compatibilism did not depend on strong reliabilist assumptions.

6. Switching cases: actual and possible.

Another way of countering Boghossian's incompatibilist argument is to hold that the mere possibility of switching cases does not prove that we lack self-knowledge. Only actual switching cases would threaten

self-knowledge. Ted Warfield (1992) has taken this line of response. He summarizes Boghossian's argument as follows (where "P" = "S's thought is about water" and "S" designates an individual in our world):

"P1. To know that P by introspection, S must be able to introspectively discriminate P from all relevant alternatives of P.

P2. S cannot introspectively discriminate water thoughts from twater thoughts.

P3. If the Switching case is actual, then twater thoughts are relevant alternatives of water thoughts.

C1. S doesn't know that P by introspection" (Warfield 1992, p. 235)

Warfield holds, correctly, that this argument is not valid. All that follows from premises P1-P3 is the much weaker conclusion C1':

C1'. If the Switching case is actual, then S doesn't know that P by introspection.

In order to obtain C1 an additional premise, P*, would be needed:

P*. The Switching case is actual.

Boghossian's argument shows, at most, that, given externalism, it is not necessary that the contents of a subject's thoughts are knowable to him on the basis of introspection.

In a different paper, Warfield gives the following informal summary of his response to Boghossian: "Boghossian argues that if externalism is true, an individual (called a Traveller) who is somehow transported back and forth between Earth and Twin Earth will not know the content of the thought she expresses with the sentence 'Water is wet'. Leaving aside the question of whether or not Travellers can have knowledge of the contents of their 'water' thoughts, I show in my (1992) that Boghossian's argument shows at most that *Travellers* do

not know the contents of their thoughts; it does not show that we do not know the contents of our thoughts because Boghossian has not argued that we are Travellers" (Warfield 1995, p. 540).

Peter Ludlow (1995a) has tried to reply to Warfield. According to Ludlow, premise (P3) in Warfield's reconstruction of Boghossian's argument is unnecessarily strong in stating the conditions under which there are relevant alternatives to water thoughts. A weaker premise like (P3') would do as well:

(P3') If switching cases in general are prevalent, then there are relevant alternatives of water thoughts.

In order to show that (P3') is sufficient, Ludlow offers the following analogy. Suppose that counterfeit is frequent at coin shows. Thus, even if there is no counterfeit at a particular coin show, a subject cannot be said to know that a certain coin is authentic if he is not able to discriminate it from a false replica. The alternative that a particular coin is false is relevant even if, as it may happen, there are no false coins in this particular context.

But if (P3') is the right premise, then a premise weaker than P* would be sufficient to establish Boghossian's incompatibility thesis, something like (P**):

(P**) Switching cases, in general, are prevalent.

He tries to show that this premise is true by construing a Burge-style case in which a subject, Biff, moves back and forth between communities largely overlapped with respect to language (namely British English and American English speaking communities). Biff uses the term 'chicory' deferring successively to each community without noticing that this term has a different meaning in each. So, when Biff has a thought involving that term in England he is thinking one thing,

but when a similar episode takes place in the United States, he is thinking something different. However, he does not know that he is having different thoughts. In order to substantiate (P**), Ludlow generalizes Biff's case by saying that "we routinely move between social groups and institutions, and in many cases shifts in the content of our thoughts will not be detected by us. (There is, it appears, a little of Biff in all of us)" (Ludlow 1995a, p. 48).

It is worth noting that, if the proposal we are going to make below is correct, externalism does not necessarily entail the consequence Ludlow wants to draw from this example. But let us grant provisionally this consequence for the sake of the argument. In my opinion, Warfield's requirement that Switching cases be actual is misguided, as, correspondingly, is Ludlow's attempt to defend Boghossian's argument by showing such cases to be prevalent. Warfield is probably right that, strictly speaking, Boghossian's argument does not establish that we actually lack self-knowledge. Only the actuality or at least the prevalence of Switching cases would have that general consequence. As Warfield points out, the right conclusion of Boghossian's argument is that, given externalism, it is possible that we lack self-knowledge. But this conclusion is devastating enough if we carefully reflect on what it involves. Since Switching situations are possible, whether or not they are actual or prevalent, and since subjects in those situations may be fully unaware that they are in them, Boghossian's argument does establish that, given externalism, self-knowledge is contingent on circumstances (such as an undetected change in our environment) we might not be aware or have any control of. Externalism entails the possibility of situations where a subject's being mistaken about the world leads to his being mistaken about what he is thinking. Owens

explicitly acknowledges this when he writes: "Because of her lack of information about the world a subject may have mistaken beliefs about her beliefs" (Owens 1995, p. 265). But this is unacceptable, whether or not we are victims of Switching cases. If externalism entails that we may lack self-knowledge owing to our being mistaken about the world, then externalism conceals the threat of an epistemological havoc, for we could not know what we believe before knowing that our beliefs about the world are true, but we could not know that our beliefs about the world are true without knowing what it is that we believe. This circle would lead to complete epistemological darkness.

7. Towards a reasonable compatibilism.

In the following sections I will try to elaborate the makings of a reasonable form of compatibilism. My strategy will be to short-circuit Boghossian's argument at a much earlier stage than do the attempts we have been reviewing. The failure of those attempts, I suspect, is due to the fact that they concede too much to the incompatibilist. I shall be questioning instead some assumptions that underlie Boghossian's argument and allow it to get off the ground.

What makes us worry about self-knowledge, I contend, is not externalism as such, that is, the general doctrine that our thought contents and concepts are individuated, in part, by external factors, but rather a particular, though widely extended construal of this doctrine. This particular construal, which might be called "causal externalism", is explicitly assumed in Boghossian's incompatibilist argument as well as (sometimes implicitly) in the attempts to counter it we have seen so far, and is illegitimately identified with externalism as such. The core of this construal, in Boghossian's presentation of it, is the principle of content fixation we referred to above. To recall, the principle is as follows, in Boghossian's own words: "... The contents of thought tokens of a given syntactic type are determined by whatever environmental property is the typical cause of the perceptions that cause and sustain tokens of that type" (Boghossian 1992, p. 19). Given this principle, a change in the typical causes of thought tokens of a given syntactic type will lead to a change in the content of those tokens. Now, since the change of typical causes may go undetected by a subject, as happens in Switching cases, the change in his thought contents will also go undetected by him, hence (given some plausible discrimination condition) he will lack self-knowledge with respect to those thoughts.

This construal of externalism cannot but fall prey to Switching cases arguments. But nothing forces us to accept this construal. In fact, I think we should reject it and turn to a different, more plausible construal.

Recall that, according to Boghossian, our intuition about the semantics of Switching cases is that, after an indeterminate period on a twin environment, tokens of 'water' in the Travellers' mentalese change their meaning, with a change in the content of the corresponding thought tokens (cf. Boghossian 1992, p. 18). I want to challenge that intuition. Intuitions, as we know, constitute a shaky, problematic domain. In many cases it is rather unclear whether they are really independent of previous theoretical commitments or are unwittingly fueled by these commitments themselves. I guess that the intuition Boghossian refers to is fed by a commitment to a causal construal of externalism, so it cannot be used to validate this construal. Another problem with this intuition is that it is raised by such extreme examples as transworld unwitting travelling cases, and it is doubtful that our conceptual background is fit enough to resist the strain and to yield clear verdicts when confronted with such extraordinary scenarios. A final point is that the intuition and the principle of content fixation that, according to Boghossian, accounts for it rest on the unwarranted attribution to environmental factors of an unexplained, unanalyzed and, let me say, sort of magical power to gradually influence and eventually change, after an indeterminate period of exposure to them, a Traveller's thought contents.

Let me dispute the intuition and the principle that supposedly explains it by resorting to more mundane, more realistic switching cases, where a factor in a subject's environment changes without the

subject's noticing the change. For the reason given, our intuitions about these cases are likely to be more reliable than those raised by extraordinary stages and can be then extended to the latter. Think of the following case. Suppose I have been longing to possess a three carats diamond and that, after strenuous financial efforts, my dream comes finally true. A wonderful diamond shines now inside a glass case in my sitting room. It is clear that the term 'diamond' in my mental attitudes and verbal expressions of them up to now means *diamond*. Unfortunately, an expert thief gets in my house and replaces my diamond by a false replica, a zircon piece, which I cannot distinguish from my longed for diamond. From then on, paraphrasing Boghossian's principle of content fixation, the environmental property that is the typical cause of the perceptions that cause and sustain my tokens of the syntactic type 'diamond' (cf. Boghossian 1992, p. 19) is the property of being a zircon piece, not the property of being a diamond. But do we really have the intuition that, after an indeterminate while, my tokens of 'diamond' have come to mean *zircon*? I definitely think we don't. My tokens of 'diamond' continue to mean *diamond* and I know they do. This Switching case has, no doubt, epistemological consequences, but these consequences do not affect self-knowledge. My belief that I possess a diamond is now false, as well as my belief that my desire to possess a diamond is now satisfied, but no doubt I know what these beliefs and desires are. If Boghossian's principle of content fixation were correct, my beliefs would now be true and my desire would be satisfied, for their content would have come to be about zircon, but this is surely wrong. My tokens of 'diamond' still mean *diamond* despite their being now typically caused by zircon.

A possible objection to this example[7] is that it differs in some important respects from the Switching examples that motivate Boghossian's argument. In particular, it might be held that, in my example, what might be keeping the meaning of 'diamond' constant is the continuity of the social community to which the subject defers and in which 'diamond' means *diamond*. In order to meet this objection, we can try to construe a case where the shift also affects the social community. One problem with this attempt is that the example becomes less realistic and the intuitions it raises might correspondingly be less reliable. Anyway, I will try to make the example as realistic as possible. Imagine, then, that the theft occurs shortly after I have moved to another country where still English is spoken, with the only difference, which I do not know about, that members of the new community use 'diamond' to mean what in the original community was meant by 'zircon' and conversely. Does this imply, given Burge's social externalism, that after living in the new community long enough the thoughts I express with tokens of 'diamond' turn into thoughts about zircon? Think that the experts of the new community, if they were to examine the mineral, would assent to my utterance "I possess a nice diamond" (though they might be slightly surprised to find that I am so proud of my zircon). Though this case is a bit more complicated, my intuition is definitely that my tokens of 'diamond' still mean *diamond*, not *zircon*. With 'diamond' I do not want to unconditionally mean whatever it is that my new community (or its experts) mean by 'diamond'; I only intend to mean that under the assumption that what they mean by 'diamond' is the same as that which is meant by that word in my original community. Since this assumption is false, I still

---

[7] I owe this objection to Andreas Kemmerling.

defer to my old community for what concerns the meaning of 'diamond'. The original community keeps its preeminence over the new one. To see this, suppose that I get to know about the semantic difference. From then on, I would certainly tell members of my new community that I possess a wonderful zircon (provided I am still ignorant of the theft), thus changing the word to preserve the meaning. And I would be sadly surprised to be told that what I possess is a diamond, not a zircon, as sadly surprised, in fact, as if in my original community I had been told that I possess a zircon and not a diamond. This speaks clearly, I think, in favour of the view that, while ignorant of the semantic difference, the meaning of my tokens of 'diamond' has not shifted, nor has the content of the thoughts I express with tokens of that word. This intuition is consistent with Burge's stress on the importance of social communities in fixing thought contents; it only requires to accept that deference to the actual community one happens to live in need not be unconditional: a subject may defer to a different community, especially if this is the community where he grew up.

My suggestion is that something similar would apply to the case of our hero Peter, the inter-world traveller. Peter's tokens of 'water' on Twin Earth continue to mean *water*, in spite of their being now typically caused by twater. In view of our modified example, it will not do to say that, after staying on Twin Earth long enough, Peter's deference to the Twin Earthian community makes his tokens of 'water' mean *twater*. Peter is only deferent to his new community under the assumption that this community is that in which he learned language and its meaning. Since this assumption is not true, Peter still defers to his original Earthian community, and his tokens of 'water' still mean *water*.

8. Normative externalism.

Let me now try to explain the intuitions about realistic switching cases. Someone might be tempted to think that the only possible theoretical explanation of them is internalism. If this were so, then compatibilism would have lost the battle. But this is not so. A plausible externalist view of meaning and content fixation can also account for them. Let me call the construal of externalism I favour "normative externalism".[8] We can contrast normative externalism and causal externalism by noticing their respective answers to the following question: Why do tokens of 'water' mean *water* on Earth and *twater* on Twin Earth? The answer of causal externalism is, roughly, as follows: because Earthians' tokens of 'water' are typically caused by water, while Twin Earthians' tokens of 'water' are typically caused by twater. It is clear how this conception leaves open the possibility of Switching cases. According to normative externalism, in turn, the answer would be: because Earthians learn and teach the meaning of 'water' in connection with paradigmatic samples of water, while Twin Earthians learn and teach the meaning of 'water' in connection with samples of a different substance, namely twater. On this construal of externalism, our words' meaning depends on external conditions because certain bits of the external world are used as samples in order to *define* those words, to give those words their meaning. The external sample becomes a norm for a correct use of the word. So, suppose that, in using certain words in thought and talk, we implicitly rely on the paradigmatic samples in connection with which we learned the meaning of those words and that we defer, unless we have positive reasons for doing

[8] Nenad Miscevic suggested to me the label "definitional externalism" for my position. His reasons will be apparent from what follows. Nothing really substantial hinges on the label, but I still prefer "normative externalism".

otherwise, to the original community where we learned that meaning. This would allow for a high degree of constancy in our words' meaning and in the thoughts we express with them, a constancy they would not have if meaning depended just on the typical causes of tokenings of those words.

It seems clear to me that this construal of externalism is consistent with Burge's social externalism. In fact, given the emphasis it puts on the social interaction involved in the process of language learning and teaching, this construal is much easier to square with Burge's insights than the causal construal. Besides, normative externalism is not really far from Putnam's conception either. Remarks about paradigmatic samples can also be found in Putnam's "The meaning of 'meaning'". Normative externalism and causal externalism, unlike internalism, can account for our intuitions about Putnam's original Twin Earth thought experiments. But normative externalism and internalism, unlike causal externalism, explain our intuitions about realistic Switching cases. So, only normative externalism can account for both groups of intuitions, which clearly speaks in its favour. It explains why we tend to judge that tokens of 'diamond' do not come to mean *zircon* after being, from a certain moment on, typically caused by zircon. The meaning of 'diamond' is not fixed by typical causes of tokenings of that word, but by paradigmatic samples of diamonds, and this is why a change in those typical causes does not affect the word's meaning.

If we extend our intuitions about realistic cases to more extreme, inter-world switching cases, something quite similar can be said about Peter. Peter learnt the meaning of 'water' on Earth in connection with samples of water, so that, with respect to this meaning, he defers to the

real nature of these samples (and to the Earthian community where they are used) in tokening the word in speech or in thought. Since, after his unwitting travelling to Twin Earth, no new process of learning takes place, his tokenings of the word retain their Earthian meaning. They still mean *water*, in spite of their being now typically caused by twater. Consequences of all this for self-knowledge should be clear by now. Peter retains his introspective knowledge of comparative content; his thought contents are, in Boghossian's terms, transparent for him. Peter's judgment is that the thought he expresses, when he is on Twin Earth, in tokening the sentence "water quenches thirst" has the same content as the thought he expressed on Earth, some time ago, by tokening that sentence. According to causal externalism, Peter is wrong on this account. But according to normative externalism, *he is right.* Some of Peter's beliefs are now false on Twin Earth, but his beliefs, even comparative, about the contents of those beliefs are still true. If all this is correct, the inclusion model of self-knowledge, completed with a normative construal of externalism, can successfully meet the Switching cases objection to compatibilism.

My proposal can temperate the crude externalist reliabilism which underlies the inclusion model, thus making the latter less vulnerable to shortcomings of the former. I think it is correct to say that self-ascriptions include the content of the first-order thought, but on my account this content does not get fixed in complete independence of the subject's intentions to keep faithful to certain content-giving practices and definitions. The real nature of water, for example, contributes to the meaning of 'water', as does in the causal version of externalism, but not outside a social interaction frame in which the subject who uses the word is knowingly involved. This allows my

proposal to prevent a mere shift in the causal (or even social) environment from automatically producing a shift in meaning and content.

It might be objected that, according to my account, if Peter's unwitting travel to Twin Earth takes place while he is still learning the meaning of 'water', so that this learning continues on Twin Earth, his term 'water' will come to mean *water-or-twater*. I accept this. But I do not think it is a problem for my account, for something similar would have happened on Earth if water had turned out to be a collection of different substances, as is the case, e. g., with jade. Peter would retain comparative self-knowledge. I conclude, then, that, at least for what concerns the Switching cases objection, externalism, on a normative reading, and self-knowledge, even in its stronger, comparative form, are compatible.

Carlos.Moya@uv.es

References

Bernecker, S. 1996. 'Externalism and the Attitudinal Component of Self-Knowledge', *Nous* 30, 262-275.

Bernecker, S. 1998. 'Self-Knowledge and Closure', in P. Ludlow & N. Martin (eds.), *Externalism and Self-Knowledge*, CSLI Press, Stanford, 333-49..

Bilgrami, A. 1992. *Belief and Meaning*, Blackwell, Oxford/Cambridge Mass.

Boghossian, P. A. 1989. 'Content and Self-Knowledge', *Philosophical Topics* 17, 5-26.

Boghossian, P. A. 1992. 'Externalism and Inference', in E. Villanueva (ed.), *Rationality in Epistemology*, Philosophical Issues 2, Ridgeview Publishing Company, Atascadero (California), 11-28.

Boghossian, P. A. 1994. 'The Transparency of Mental Content', *Philosophical Perspectives* 8, 33-50.

Boghossian, P. A. 1997. 'What an Externalist Can Know A Priori', *Proceedings of the Aristotelian Society* 97, 161-75. Reprinted in E. Villanueva (ed.), *Concepts, Philosophical Issues* 9, 1998.

Bonjour, L. 1992. Entry 'Externalism/internalism', in J. Dancy and E. Sosa (eds.), *A Companion to Epistemology*, Blackwell, Oxford, 133-136.

Brown, J. 1995. 'The Incompatibility of Anti-individualism and Privileged Access', *Analysis* 55, 149-156.

Brown, J. 1999. 'Boghossian on Externalism and Privileged Access', *Analysis* 59, 52-59.

Brueckner, A. 1997. 'Externalism and Memory', *Pacific Philosophical Quarterly* 78, 1-12.

Brueckner, A. 2000. 'Externalism and the A Prioricity of Self-Knowledge', *Analysis* 60, 132-136.

Burge, T. 1988. 'Individualism and Self-Knowledge', *Journal of Philosophy* 85, 649-663.

Burge, T. 1993. 'Content Preservation', *The Philosophical Review* 102, 457-488.

Burge, T. 1996. 'Our Entitlement to Self-Knowledge', *Proceedings of the Aristotelian Society* 91, 91-116.

Burge, T. 1998. 'Memory and Self-Knowledge', in P. Ludlow and N. Martin (eds.), *Externalism and Self-Knowledge*, CSLI Press, Stanford, 351-70.

Davidson, D. unpubl. ms. Quoted in Boghossian 1994, p. 35.

Falvey, K. 2000. 'The Compatibility of Anti-Individualism and Privileged Access', *Analysis* 60, 137-42.

Falvey, K. and Owens, J. 1994. 'Externalism, Self-Knowledge, and Skepticism', *The Philosophical Review* 103, 107-137.

Gibbons, J. 1996. 'Externalism and Knowledge of Content', *The Philosophical Review* 105, 287-310.

Goldberg, S. C. 1997. 'Self-Ascription, Self-Knowledge, and the Memory Argument', *Analysis* 57, 211-219.

Heil, J. 1988. 'Privileged Access', *Mind* 97, 238-251.

Heil, J. 1992. *The Nature of True Minds*, Cambridge University Press, Cambridge.

Ludlow, P. 1995a. 'Externalism, Self-Knowledge, and the Prevalence of Slow Switching', *Analysis* 55, 45-49.

Ludlow, P. 1995b. 'Social Externalism, Self-Knowledge, and Memory', *Analysis* 55 (1995), 157-159.

Ludlow, P. and Martin, N. (eds.) 1998. *Externalism and Self-Knowledge*, CSLI Press, Stanford.

McKinsey, M. 1991. 'Anti-Individualism and Privileged Access', *Analysis* 51, 9-16.

Moya, C. J. 1998. 'Boghossian's Reductio of Compatibilism', in E. Villanueva (ed.), *Concepts. Philosophical Issues* 9, 243-251.

Moya, C. J. 1999. 'Self-Knowledge and Content Externalism', in J. Nida-Rümelin (ed.), *Rationality, Realism, Revision*, W. de Gruyter, Berlin/New York, 182-187.

Owens, J. 1995. 'Pierre and the Fundamental Assumption', *Mind and Language* 10, 250-273.

Warfield, T. 1992. 'Privileged Self-Knowledge and Externalism Are Compatible', *Analysis* 52, 232-237.

Warfield, T. 1995. 'Knowing the World and Knowing Our Minds', *Philosophy and Phenomenological Research* 55, 525-545.

Woodfield, A. 1982. 'Foreword' to *Thought and Object. Essays on Intentionality*, Clarendon Press, Oxford, v-xi.