(Pre-print)

# SELF-KNOWLEDGE AND CONTENT EXTERNALISM[1]

Carlos J. Moya. University of Valencia

The question whether direct, authoritative knowledge of one's own thought contents and an externalist individuation of those contents are compatible has been, and still is, the object of a wide discussion. In this paper I shall present the main lines of the discussion and put forward the makings of an affirmative answer to the compatibility question. Owing to space restrictions, my presentation is bound to be rather sketchy, though I will try to bring out the central points of my perspective as clearly as possible.

The simplest way in which incompatibilism could be established would be to start from the premise according to which if content depends on external factors, knowledge of content must depend on knowledge of those factors (cf. Bonjour 1992, p. 136). Given externalism, this premise would entail that in order to know what we think we should first investigate our surroundings, which in turn leads to the conclusion that we do not have direct, authoritative knowledge of our own thoughts. This premise, however, does not seem to be true. Think, for example, that though my existence (metaphysically) depends on my parents' existence, I can know that I exist even if I do not know about my parents' existence. This holds not only in cases of metaphysical dependence, but also in cases of conceptual dependence: someone can know that a certain figure is a triangle and not know that its internal angles measure 180 degress even though this figure's being a triangle depends upon its internal angles' measuring 180 degrees.

Some philosophers (Burge 1988, Heil 1988, Davidson, unpubl. ms) have tried to defend compatibilism by noting that reflexive self-ascriptions of thoughts include the content of the ascribed thought itself, whatever the way this content is determined. On the inclusion model of self-knowledge, as this proposal might be called (Bernecker 1996), Cogito-like judgments are reliably true in that they are contextually self-verifying, as Burge insists. A subject need not know what the individuation conditions of his thoughts are in order to correctly ascribe these thoughts, with their right contents, to himself.

One major objection to the inclusion model has been put forward by Boghossian (1989, 1992), on the basis of thought experiments in which a subject is unwittingly switched between distinct but observationally undistinguishable environments, say between Earth and Twin Earth. Let's baptize our inter-world traveller 'Peter'. Suppose that Peter is unwittingly

[1] 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 want also to express my gratitude to Carlos Ulises Moulines, Julian Nida-Rümelin and Wilhelm Vossenkuhl for inviting me to present a version of this paper to the 3rd Congress of the *Gesellschaft für Analytische Philosophie*. I am also grateful to Sven Bernecker, Andreas Kemmerling and Nenad Miscevic for their useful comments and criticism.

transported to Twin Earth. 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* [*twin water*, C. M.]" (Boghossian 1992, p. 18). This intuition can be accounted for by the following principle of content determination, which, according to Boghossian, is at the basis of 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). The consequence Boghossian wants to draw from this thought experiment is that Peter lacks comparative knowledge of his thought contents. He cannot discriminate between the thought that water is a good drink (a thought he expresses with "water is a good drink") and the thought that twater is a good drink (a thought he expresses with those same words). On this basis, Boghossian thinks he can show that Burge's Cogito-like, self-verifying judgments do not constitute knowledge. Suppose, in effect, that, after being on Twin Earth long enough, Peter is told that he has been switched, but not when the switch took place. Now, Peter will not know, in the circumstances, whether (say) two years ago he was thinking that water is a good drink or that twater is a good drink. Boghossian, however, takes the following to be a platitude about memory and knowledge: "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). If we stipulate (as seems reasonable) that Peter suffers no memory failure, this shows that he did not know, two years ago, what thought he was expressing with "water is a good drink", even if the self-ascription expressed by the sentence "I judge that water is a good drink" happened to be true in virtue of the content inclusion mechanism. Boghossian's conclusion is that "Burge's self-verifying judgments do not constitute genuine knowledge" (Boghossian 1989, p. 23).

Attempts to refute Boghossian's incompatibilist argument have moved along (at least) three different lines but, to my lights, none of these lines is clearly successful.

According to the first line (cf. Falvey and Owens 1994), we do not enjoy, independently of externalism, direct comparative knowledge of our thought contents. The price this line has to pay is to reject a seemingly correct necessary condition for knowledge, namely the ability to discriminate between relevant alternatives. It is plausible to think that someone's judgment that *a* is an *F*, even if this judgment is true, does not amount to knowledge of that fact if, in case *a* were a *G*, where being a *G* is a relevant alternative to being an *F*, he still would judge that *a* is an *F*. Note that what counts for possession of knowledge is the existence of relevant alternatives, not the subject's belief or knowledge of this existence. Now, Peter, the unfortunate inter-world traveller, does not satisfy this condition, even if his judgments about his thought contents happen to be true (and reliably so).

A second line of response focuses on the role memory plays in Boghossian's argument. Goldberg (1997) and Brueckner (1997) have disputed Boghossian's supposed platitude about memory and knowledge. According to these authors, even if S forgets nothing, he may know that p at a certain time and not know that p at a later time, owing to the fact that, meanwhile, he has acquired new information which defeats his justification for believing that p. This is precisely what happens to Peter when he is told about the switch. However, this objection, correct as it may be, is not decisive, for Boghossian's argument can be restated without any essential appeal to memory. Suppose, in effect, that Peter is told that he has been repeatedly switched between Earth and Twin Earth, but not when the switches occurred nor where he finds himself at this moment. He will acknowledge that he does not know, right now, whether he is thinking, right now, that water is a good drink or that twater is a good drink. In fact, as we noted in connection with the relevant alternatives condition for knowledge, it is not even necessary, in order for Peter's self-knowledge to be put in question, that he be *told* about the switches. If the switches have in fact taken place, he lacks comparative self-knowledge, whether he knows about them or not, for he cannot discriminate between relevant alternative contents.

A third line of response (Warfield 1992) insists that the mere possibility of switching cases does not establish that we lack self-knowledge. Only actual switching would have this consequence. The most that Boghossian has shown is that, given externalism, we might lack self-knowledge. He has not shown that we actually lack self-knowledge because he has not argued that we are actually switched. Ludlow (1995) has tried to meet Warfield's objection by arguing that switching cases are quite frequent and that this is enough to run a Boghossian-style incompatibilist argument. On the basis of Burge's social externalism, Ludlow insists that, since we defer to social communities with respect to the meanings of our words and we often move between social groups and institutions with different semantic rules for the same words (think, e. g., of the different meaning of "realist" in philosophical and non-philosophical circles), the contents of our thoughts often shift in ways that are not detectable by us. Now, even if Ludlow is wrong, Warfield's objection does not seem to me powerful enough to undermine incompatibilism. Warfield may be right that, strictly speaking, the conclusion of Boghossian's argument is not that, given externalism, we actually lack self-knowledge, but only that we might lack it. But this conclusion is strong enough if we carefully reflect on its import. If this conclusion is true, then externalism entails the possibility of cases where a subject is wrong about his thought contents in virtue of his being wrong about the external world. Owens explicitly grants 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). If so, however, externalism has unacceptable epistemological consequences, for, in order to know what we believe, we should first ensure that our beliefs about the world are true, but we could not do

this without first knowing what these beliefs are. This circle would amount to an epistemological collapse.

None of the compatibilist rejoinders I have been reviewing so far is indisputably successful against switching cases arguments. The common reason for their different shortcomings, I suggest, is that all of them concede too much to the incompatibilist. All these attempts agree (at least implicitly) with Boghossian that the thought contents of an inter-world traveller shift with the change in the environment that is causally responsible for these thoughts. So, they endorse the externalist principle of content determination stated by Boghossian and quoted above. I strongly suspect that, once this principle is accepted, there is no way for an externalist to avoid falling prey to switching cases objections. But is an externalist committed to accepting this principle and its consequence that a shift in the causal enviroment leads to a shift in thought contents? I do not think so. Externalism is the doctrine that concepts and thought contents are partly determined by external factors. But the above principle of content determination is only one particular way in which this general doctrine can be spelled out and should not be confused with externalism as such. Let me call this particular spelling "causal externalism", given the centrality of causal relations in fixing meanings and intentional contents. I shall try to show that causal externalism is not true.

Do we really have, independently of a theoretical commitment to causal externalism, the intuition that meanings and mental contents expressed by certain linguistic tokens change with a change in the typical causes of those tokens? I doubt this. Let us focus on more realistic cases of switching, where an undetected change in a subject's environment occurs. Since our concepts are put under less pressure, intuitions prompted by these cases are likely to be more reliable than those raised by such extreme scenarios as inter-world travels. Think of the following example. Suppose some (or all) pieces of metal in the house of an old rich man (let us call him 'Robert') are made of platinum. Robert never gets out, so his only causal contact with platinum reduces to the pieces of his house. It is clear, I hope, that 'platinum', in Robert's mouth and inner thoughts, expresses the concept *platinum*. He has several attitudes involving this concept. He believes, for instance, that he has a lot of platinum at home, that platinum is an expensive and valuable metal, and so on. Now imagine that, while Robert is sleeping, an expert thief gets in the house and replaces all pieces of platinum by similar pieces of chromium-plating. Suppose that chromium-plating is observationally undistinguishable from platinum and that Robert never discovers the change. So, from then on, to paraphrase Boghossian's (and causal externalism's) principle of content determination, the environmental property that is the typical cause of the perceptions that cause and sustain Robert's tokens of the syntactic type 'platinum' is the property of being chromium-plating, not the property of being platinum. But I, for one, do not have the slightest intuition that, after an indeterminate while, tokens of 'platinum', in Robert's mouth, come to mean *chromium-plating* and that, from a certain moment

on, he starts believing that he has a lot of chromium-plating (instead of platinum) at home. 'Platinum' continues to mean *platinum*. Robert still believes (though now falsely) that he has a lot of platinum at home and he knows he believes this. He now has some false beliefs about the world, but his beliefs about these beliefs' contents are true. Robert keeps his self-knowledge in spite of the switching.

Suppose that, as our intuitions about this case suggest, thought contents do not necessarily shift with a change in their typical causes. One might be tempted to think that only internalism can explain this fact, so that externalism would be false. And this would vindicate incompatibilism once again. I agree that causal externalism cannot explain this intuition, but I think that other plausible versions of externalism can account for it. Suppose, in effect, that some of our words (including natural kind terms such as 'water') are given their meanings and connected to the world by means of (ostensive) definitions that link the word to a (paradigmatic) sample of the corresponding substance. I do not think this is contentious. As Davidson says: "It is plain that we learn what many simple sentences, and the terms in them, mean through ostension" (Davidson 1994, p. 233). Though causal relations are involved in this process of teaching and learning through ostensive definitions, these causal relations are not given the last word in fixing the meaning: if, by mistake, we use a sample of alcohol to teach the meaning of 'water', we implicitly assume that this definition fails to give the right meaning of 'water'. I take it that this view of the process of meaning fixation is externalist, for we leave the real nature of the right samples to fix the meaning, even if we are ignorant of that nature. I am aware that the picture is very crude and that it should be refined and completed so as to face some problems, but it is hard to deny that something of that sort must be centrally present in the process of explaining and learning the meaning of some important parts of language.

Now 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 otherwise, to the original community where we learned that meaning. This would give our words' meaning a constancy (or robustness, to use a Fodorian term) they could not have if their meaning happened to depend just on the typical causes of tokens of those words: it would prevent this meaning from changing because of an undetected shift (as it occurs in switching situations) in those typical causes and, thereby, would protect, so to speak, the thought tokens expressed with tokens of those words against a corresponding change in their content. Let me call this view of meaning and content determination 'normative externalism'.

Let me defend normative externalism against both internalism and causal externalism by appealing to the respective explanatory capacity of these positions. The explananda are, first, the original intuitions raised by standard Twin Earth thought experiments, namely that words like 'water' do not mean the same on Earth and on Twin Earth and that the thought tokens

respectively expressed by 'water is a good drink' differ in content; and, second, the intuitions raised by realistic switching cases, namely that the meaning of certain words does not change even if there is a change in the typical environmental causes of tokenings of those words, as happens in our example. Now, this is my defence of normative externalism: internalism can account for the second explanandum, but not for the first; causal externalism, in turn, can account for the first but not for the second; but normative externalism can account for both. Though this is not a conclusive argument for normative externalism, it certainly increases the plausibility of this position, i. e., the probability of its truth.

Causal externalists may reply that normative externalism, unlike causal externalism, does not account for intuitions raised by extreme, non-realistic switching cases, such as examples of unwitting inter-world travelling. My response, as I anticipated, is that we do not have clear intuitions about those cases, for their extraordinary character makes our concepts unfit to yield definite verdicts about them. So, we should not treat those supposed intuitions as genuine explananda, for it is doubtful that they state facts.

We should better view extreme, non-realistic switching cases in the light of our reactions to, and intuitions about, realistic ones: in the same way in which tokens of 'platinum' in Robert's mouth do not come to mean *chromium-plating* after being caused by chromium-plating, tokens of 'water' in the mouth of Peter, the inter-world traveller, do not come to mean *twater* after being caused by twater; 'platinum' still means *platinum* and 'water' still means *water*. Normative externalism can explain this. Meaning is not fixed by typical causes of tokenings of words, but by the right samples used to define, to teach and to learn those words' meaning. Robert learnt the meaning of 'platinum' in connection with right samples of platinum and Peter learnt the meaning of 'water' in connection with right samples of water. They implicitly rely on those samples (even if they are ignorant of their micro-structural properties) in using those words. Moreover, both of them defer to the community where they learnt those words' meaning, for they do not have positive reasons for not doing so. In Peter's case, this means that 'water' retains its Earthian interpretation in spite of the switching, for no new process of ostensive learning has taken place after his unwitting switching to Twin Earth, nor has he reasons for not deferring to the Earthian community (which he takes to be identical to the community he now lives in). Consequences of all this for self-knowledge are now clear. Peter retains his comparative self-knowledge in spite of the switching: his judgment will be that the thought tokens he expresses, either on Earth or on Twin Earth, with "water is a good drink" have the same content. But if normative externalism is correct, this comparative judgment of Peter's about his thought contents is true: the two thought contents *are* of the same type. If this is on the right lines, the inclusion model of self-knowledge, coupled with a normative construal of externalism, can successfully meet the Switching cases objection to compatibilism. So, externalism and self-knowledge are compatible.

*E-mail: Carlos.Moya@uv.es*

*References*

- Bernecker, S., 1996, Externalism and the Attitudinal Component of Self-Knowledge. *Nous* 30, 262-275.
- 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, Atascadero, 11-28.
- Bonjour, L., 1992, entry Externalism/Internalism, in J. Dancy and E. Sosa (eds.), *A Companion to Epistemology*, Oxford, Blackwell.
- Brueckner, A., 1997, Externalism and Memory. *Pacific Philosophical Quarterly* 78, 1-12.
- Burge, T., 1988, Individualism and Self-Knowledge. *Journal of Philosophy* 85, 649-663.
- Davidson, D., unpubl. ms. Quoted in Boghossian, P. A., The Transparency of Mental Content. *Philosophical Perspectives* 8 (1994), 33-50.
- Davidson, D., 1994, entry Davidson, Donald, in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Oxford, Blackwell.
- Falvey, K. and Owens, J., 1994, Externalism, Self-Knowledge, and Skepticism. *The Philosophical Review* 103, 107-137.
- Goldberg, C. G., 1997, Self-ascription, self-knowledge, and the memory argument. *Analysis* 57, 210-219.
- Heil, J., 1988, Privileged Access. *Mind* 97, 238-251.
- Ludlow, P., 1995 Externalism, Self-Knowledge, and the Prevalence of Slow Switching. *Analysis* 55, 45-49.
- 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.