WHAT IS EXTERNALISM?

Katalin Farkas
farkask@ceu.hu

Philosophical Studies 112/3 (February 2003): 187-208

Abstract
The content of the externalist thesis about the mind depends crucially on how we define the distinction between the "internal" and the "external". According to the usual understanding, the boundary between the internal and the external is the skull or the skin of the subject. In this paper I argue that the usual understanding is inadequate, and that only the new understanding of the external/internal distinction I suggest helps us to understand the issue of the compatibility of externalism and privileged access.

Externalism about cognitive content has been discussed for almost forty years, and became almost an orthodoxy in the philosophy of mind. This orthodoxy assumes a general rough and ready understanding of the externalist thesis, without there being an unanimous agreement over its precise nature. Details of an exact definition perhaps do not matter for certain purposes, but they are important if we want to draw further consequences of the doctrine; for example in considering the compatibility of externalism and self-knowledge. This debate has reached an almost hopelessly labyrinthine state, and the reason lies, I think, partly in a certain confusion about what externalism is. In what follows, I shall try to clarify this issue.

1. The boundary between the external and the internal

A number of views have been called "externalist" even within the philosophy of mind. I cannot hope to discuss all of them here, so I shall focus my attention on what may be called "Twin Earth externalism", the version of externalism which is expressly motivated by Twin Earth style arguments. Twin Earth style arguments aim to establish - through the analysis of concrete examples - that the following is possible: that two subjects should have qualitatively identical internal states and yet the content of (some of) their mental states would be different because of some difference in their external environment.[1] The prototype of these arguments is, of course, Putnam's argument in "The Meaning of 'Meaning'". These arguments are used to support the thesis of externalism, which can be formulated for example as follows:

[1] Some philosophers use the Twin Earth strategy to define externalism. See for example McLaughlin and Tye 1998, 285 and in Davies 1998, 327. Jackson and Pettit draw the distinction between "narrow" and "broad" content in terms of the notion of a *Doppelgänger*; see Jackson and Pettit 1996, 220.

- the content of a subject's thoughts depends on or is individuated by facts *external* to the subject; or that
- the content of a subject's thoughts does not supervene on her *internal* states; or that
- a subject's having certain thoughts presupposes the existence or particular nature of things that are *external* to the subject.

There may be other versions, but all versions agree in one point, namely drawing a boundary between the external and the internal or some related notions. These formulations certainly capture at least part of what externalism is, but they will be incomplete without answering a crucial question: what do the phrases 'external' and 'internal' mean? How should we draw the boundary between the internal and the external?

There is one interpretation which seems to be accepted in many discussions: that 'external' means '*external to the body or skin (or brain) of the subject*'[2]. Then the externalist thesis claims that the content of a subject's thoughts or sentences depends on facts external to her skin. This conception certainly gets support from Putnam's original formulation, that 'meanings ain't in the head'. However, I shall try to show that the point of externalism is not really about the individuating facts being inside or outside the skin.

As I said, my interest here lies in the version of externalism expressly motivated by Twin Earth style arguments, which feature two subjects whose internal states are stipulated to be the same. This suggests a way of finding out what 'external' and 'internal' mean. We have to focus on the *relation between the Twins*; what is this thing they share and which, according to the externalist, is not sufficient to individuate the content of their thoughts? If we can say what this is, then we have a grasp on what 'internal' is; and everything which the Twins may not share will be 'external'.

Given the assumption that 'internal' means 'inside the skin' and 'external' means 'outside the skin', the usual strategy has two subjects whose in-the-skin states are (qualitatively) physically identical. So the relation between the Twins is *identity in qualitative physical make-up*. In what follows, I argue that the stipulation about identity in physical make-up is neither necessary, nor sufficient for the externalist argument to proceed. This means that the interpretation of 'internal' as 'inside the skin' is inadequate; the boundary between the internal and the external should not be drawn around the skin.

2. The stipulation of identity in physical make-up is not sufficient

If an argument succeeds in showing that two subjects, whose physical make-up is identical, nonetheless have different mental contents, this conclusion will rule out a number of theories of

[2] Examples are many; see for example Davies 1998, 322; McLaughlin and Tye 1998, 285; MacDonald 1998, 124, Boghossian 1997, 163. Most authors referred to in this paper make similar assumptions, see McCulloch 1995, 189, Jackson and Pettit 1996, 220; Burge 1988, 650

mind. Since brain states, functional states and behavioral dispositions (on a certain construal) supervene on bodily states, outside-the-skin externalism can be used to refute for example the identity theory, functionalism and (certain versions of) behaviorism.[3] Such an argument, however, fails to address dualist versions of internalism. To oppose dualism in an externalist spirit, we should present two subjects whose mental states were identical *according to the dualist's criteria*, and *then* argue that their thoughts are different due to some difference in their environment. But stipulating qualitative identity in the subjects' physical make-up will not be sufficient to assure identity of mental states on the dualist conception; states of immaterial souls or non-physical properties of mental states need not supervene on bodily states.

Given a wide acceptance of materialism, addressing dualism will perhaps not be regarded as an important issue. The point, however, is not so much polemical, but explanatory. The Cartesian theory of mind is often regarded as the arch-source of internalism; but the reason to regard Descartes as an internalist cannot be that on his theory, mental states are individuated by bodily states. So the usual understanding leaves it unexplained in what sense Cartesianism is an internalist theory.[4]

In fact, Putnam himself did think that his externalism had to rule out dualism. When describing the Twin Earth scenario, he says "suppose I have a Doppelgänger on Twin Earth who is molecule for molecule 'identical' with me ... If you are a dualist, then suppose my Doppelgänger thinks the same verbalised thoughts I do, has the same sense-data, the same dispositions etc." (Putnam 1975, 227). At another point the Twins are said to have the same beliefs, thoughts, feelings and so on (ibid. 224). The problem with this is familiar: Putnam's original argument was meant to support externalism about *meanings*, but then externalism was extended to *mental content*. If the point of the Twin Earth argument is to show that the content of the Twins' beliefs are different, then you cannot set up the argument by saying that they have the same beliefs or thoughts. So whereas Putnam could perhaps help himself to the relation 'having the same thoughts' when arguing for semantic externalism, the same formulation cannot be used in an argument for externalism about mental content.[5]

We cannot characterize the relation between the Twins as 'having the same thoughts'. But if the argument requires only that the Twins should be molecule for molecule identical, then it fails to address dualism. Hence the stipulation of identity in physical make-up is not sufficient for running a general externalist argument.

---

[3] See McCulloch 1995, 168.

[4] There are philosophers who appear to acknowledge this point and try to define externalism in a way which is applicable to immaterialist theories. See for example Burge 1986, 118-9; Pettit 1986, 17-18. Gabriel Segal distinguishes between 'intrinsic' and 'locally supervenient' properties, and defines internalism primarily in terms of the former. (Segal 2000 9ff.) I think, however, that for the reasons to be spelled out in the next section, the notion of 'intrinsic' is still not suitable for defining internalism.

[5] See Burge 1982.

3. The stipulation of identity in physical make-up is not necessary

Arguably, at least some diseases are natural kinds: they have some superficial properties (the symptoms) on the basis of which we normally identify them, and some underlying structure which is responsible for the superficial properties - for example a certain inflammation caused by some bacteria. We can then design the following Twin Earth case: suppose that the disease known as 'meningitis' on Twin Earth, which has exactly the same symptoms and overall effects as meningitis on Earth, is in fact not caused by the bacterium Meningococcus (as on Earth), but by a different bacterium which we will call 'XYZ'. Consider $Oscar_1$ on Earth, suffering from meningitis back in 1750, when the bacterium causing meningitis was unknown, and $Oscar_2$ on Twin Earth, who is as similar to $Oscar_1$ as possible, except that at the same time he is suffering from the disease caused by XYZ. Then it seems that the argument could proceed in the same way as in the original Twin Earth case: $Oscar_1$'s thought 'meningitis is a dangerous disease' has a different content from $Oscar_2$'s parallel thought, because $Oscar_1$ is thinking about meningitis, and $Oscar_2$ is thinking about the disease caused by XYZ. We reached a conclusion very similar in spirit to the externalist thesis, but the relevant individuating facts in this case appear to be inside the body.[6]

It may be noticed that the original Twin Earth example involving 'water' is in fact similar to the meningitis case: as it has been repeatedly pointed out, the Twins of the original example cannot be physically identical, since the human body contains a significant amount of water. The objection is usually not regarded as very serious; the general feeling was that we could find a better example, about a substance which is not to be found in the human body, so everyone continued using the water example. I suggest that our willingness to overlook this problem in the original argument is better explained by the fact that the point of externalism is not whether the individuating facts are inside or outside the body. In fact, this becomes even clearer in a later version of the argument Putnam offers: in that version, Twin Earth water is 20 percent grain alcohol, and the body chemistry of Twin people is changed so that they react to this mixture as we do to water (Putnam 1981, 23). This argument seems to pass as an argument for Twin Earth externalism, and yet the condition of internal physical sameness is obviously violated. The meningitis example helps to bring out the point even more clearly, since our stereotype about meningitis is formed on the basis of its occurrences in the human body, whereas the same is not true of water.

What motivates the externalist analysis of the water-argument? Some philosophers refer to simple intuitions, others back up their intuitions with a certain theory of natural kind terms. It seems that whichever motivation is at work in the original example, it is also present in the

[6] I assuming here that the causal argument is *independent* from other, Burgean types of arguments for social externalism and from arguments based on the division of linguistic labour.

meningitis case. So if any argument for externalism based on natural kind terms is worth anything, the meningitis case is just as good an example as any other natural kind. Or to put it in another way: if someone claimed that $Oscar_1$ and $Oscar_2$ meant something different by 'meningitis' because of the unknown microscopic difference in their bodies, then this would be as unacceptable to someone with internalist inclinations as any other externalist conclusion. If this is right, then we have a perfectly good argument for externalism where the stipulation that the Twins would be identical in physical make-up is not necessary.

It strikes me as obvious that the point of the meningitis argument is exactly the same as the point of the water or aluminum argument, but some will perhaps disagree. The objection could run like this: "no-one denies that some in-the-skin facts make a difference to the content of our thoughts, therefore it should not come as a surprise that subjects with different physical make-up have different thoughts. The point of the Twin Earth arguments is that *even if* you stipulate molecule for molecule identity, the thoughts could still differ. Given that we have this *stronger* thesis, why should we care about the weaker thesis? Anyway, externalism *is* the view that mental contents do not supervene on bodily states, so the issue between the internalist and the externalist is whether *molecule for molecule identical Twins* can have different thoughts or not. The protagonists of the meningitis example *are not Twins in this sense*, so whatever we say about their thoughts will be irrelevant to the issue of externalism".

Let me offer the following analogy to illustrate what seems to me wrong with this objection. Suppose that we stipulate that our Twins, $Oscar_1$ and $Oscar_2$ should be molecule for molecule identical, and furthermore, that they should wear 'identical' neckties. Then we run the usual Twin Earth argument, and come up with the following interesting thesis: mental contents do not supervene on bodily states plus necktie-states. Interesting indeed, someone will say, but couldn't you run the argument without the neckties? Our reply comes readily: if we have the stronger thesis about bodily *plus* necktie states, why should we care about the weaker thesis? Anyway, necktie-externalism *is* about Twins who satisfy the given stipulations; so whatever you say about Twins who are *merely* molecule for molecule identical, it will be irrelevant to our purposes.

I assume that my opponents will acknowledge that defending necktie-externalism is pointless, but they may still remain unconvinced. They will now admit that the Twins do not have to be molecule for molecule identical: for example, $Oscar_1$ could be an inch taller than $Oscar_2$, and the argument will work just as well. But while we could all agree about particular cases, they will insist that there is no way to *specify what counts in general* as a relevant difference in bodily states - relevant in the sense that it gives the externalist his point -, and what doesn't. Neckties come off easily, but bodily parts don't. Therefore, the objection continues, the only logical or natural way to draw the boundary between external and internal is around the body. But this objection works only if there is no other way indeed to draw the boundary between the internal and the external, and the

very task of this paper is to show the contrary.

4. Subjective indistinguishability

Let me sum up the two preceding sections. First, I assumed that the externalism/internalism debate is orthogonal to the materialism/dualism debate. Therefore it seems that a general externalist thesis should be effective against dualist versions of internalism as well. Second, I argued that externalism - or something very much like it - can arise with respect to facts inside the body. If this is right, then we can ask whether it is possible to define externalism in a way which accommodates these two points. Clearly, the new definition must depart from the idea that the skin is the boundary between the internal and the external. I admit that this creates a certain a difficulty: if the usual understanding is based on the in-the-skin/outside-the-skin conception, then it seems that I simply change the subject if I propose a modification. After all, if numerous philosophers explicitly say - and they do - what they mean by externalism, then we should take their word for it. I don't want to dispute this. Still, I think it is legitimate to ask what further motives may lie behind the externalist thesis, which may be brought to the surface by considering the points about dualism and the meningitis case. I shall try to explain how my proposal overlaps with the usual understanding, and leave to the reader to decide between the usual understanding and my suggestion. I shall also argue that this choice will have significant consequences in the debate about the compatibility of externalism and self-knowledge.

So suppose that at least it makes sense to ask whether some understanding of externalism can accommodate the points raised in sections 2 and 3. What we need then is a characterization of a relation between the Twins which (i) establishes an identity or equivalence between the Twins in some respect (ii) makes the Twins in the meningitis case equivalent in this respect; (iii) implies 'internal' sameness in a way which is applicable to dualist theories. As I said earlier, the characterization of this relation will deliver the notion of the relevant boundary between the internal and the external; whatever is common to the Twins in this respect is internal to a subject, and the externalist conclusion is that this is what is insufficient to individuate mental content.

A good way to start our search for the proper characterization is to consider possible ways to *spoil* Twin Earth arguments for externalism. It is crucial in Putnam's thought experiment that water (H_2O) and twater (XYZ) should be *indistinguishable* for perceivers in normal perceptual circumstances (see Putnam 1975, 223). If twater was blue and bitter in ordinary circumstances, an internalist could easily agree that 'water' meant something different on Earth and Twin Earth. Imagine that you see now a glass of water, you taste it, it quenches your thirst. We know that the situation of Earth and Twin Earth is exactly symmetrical (they call *our* planet Twin Earth); now if *this* were Twin Earth with XYZ on it, the liquid called 'water' would look, smell and taste the same, and it would quench your thirst just as well. (In the versions where the body-chemistry varies from

Earth to Twin Earth you should imagine a counterfactual swap of body-chemistries.)

The Twin Earth argument would not work if water and twater felt different; and it would also not work if the stereotypes of water and twater were different. Suppose that the Twins had some knowledge of chemistry; if it was part of $Oscar_1$'s conception of water that it is composed of H_2O, and part of $Oscar_2$'s conception of water that it is composed of XYZ (in the sense that if they were asked about what water was, their answer would include these), then again an internalist could easily agree that 'water' meant something different on Earth and Twin Earth. So in order to make the argument for externalism work, we have to exclude such divergences; and this is, I believe, what Putnam tried to capture by saying that $Oscar_1$ and $Oscar_2$ have the same beliefs, thoughts, etc. (Putnam 1975, 224). But as I said earlier, when the issue is externalism about mental content, we cannot use the stipulation that the Twins have the same thoughts or beliefs.

How can we capture these two points? Let us summon some help. Burge is one of the few philosophers who does discuss the problem of formulating an individualist position which is applicable to non-materialist theories. His initial suggestion is that individualism is "the thesis that a person's phenomenological, qualitative mental phenomena fix all the person's mental states, including those (like thoughts, desires, intentions) with intentionality and representational characteristics" (Burge 1986, 117). This characterization seems to fit the aspect of the Twin Earth scenario I just pointed out: that water and twater should feel the same and should be associated with the same features for the Twins. To use Burge's formulation, the Twins should have the same 'phenomenological, qualitative mental phenomena' when experiencing water or thinking about water.[7]

However, Burge perceives a problem: this characterization presupposes a well-understood distinction between *phenomenological* and *intentional* aspects of the mental, and Burge thinks it's rather doubtful that Descartes and the non-materialist tradition (who were supposed to hold this version of individualism) were aware of, or would have accepted this distinction. Fortunately, we can capture the same idea without relying on the heavy conceptual machinery this distinction involves. The key is: if you were (actually or counterfactually) swapped with your Twin Earth counterpart, *things would appear the same*. This is how being transported to Twin Earth (unawares, overnight) is different from being transported say to Mars. In the latter case, the next morning things would surely look different. This idea seems to me constitutive of a proper Twin Earth scenario: that your situation is subjectively indistinguishable from your Doppelgänger's situation. But I should not like to follow Burge in thinking that this idea presupposes a distinction between the qualitative and the intentional. Suppose that someone did not agree with the compartmentalization of mental phenomena to the qualitative and the intentional, and held that all mental phenomena are

[7] I assume here, as a number of philosophers do, that all conscious thinking has phenomenological character.

intentional. It would be very strange if this philosopher could not conceptualize the Twin Earth scenario the way I put it; if she could not understand the difference between being transported to Twin Earth and being transported to Mars. Yet she would not explain this by saying that the two subjects' qualitative (as opposed to intentional) mental phenomena are the same: she could not, since she would deny that such things exist.

The notion of subjective indistinguishability is fundamental in understanding the nature of human experience, and it is prior to the qualitative/intentional distinction, or to the outcome of the externalism/internalism debate. To illustrate the latter point, consider for example the disjunctive theory of perception. Disjunctivists hold that there is no single kind of mental state a subject is in both when she is having a veridical perception or the corresponding perfect hallucination.[8] Disjunctivists are externalists, for what makes the difference between the two mental states in kind is something external to the subject. But in order to formulate the theory, we should have a grasp on what a perfect hallucination is, and this is given precisely in terms of subjective indistinguishability. A Twin scenario suggested by the disjunctivist might be something like this: now you are seeing a glass of water; but you could be hallucinating, that is, be in a situation which is subjectively indistinguishable from this one, and yet the glass is not there. So disjunctivists would accept the characterization of the relation as subjective indistinguishability; however, they would deny that it implies identity of mental states. And this is where I think internalists and externalists part: an internalist would find it difficult to accept that something which makes no difference to how a situation appears to the subject (the presence of the glass in the case of veridical perception, and its absence in the case of the corresponding perfect hallucination) could make a difference to her mental states.[9]

This is my suggestion then: the relation between the Twins is subjective indistinguishability of their situations. To repeat, this means that if they would be swapped (actually or counterfactually), things would look, feel etc, the same. This is also the same relation that holds between a veridical perception and the corresponding hallucination. And the lesson of the Twin Earth argument for externalism is that two subjects who are in subjectively indistinguishable situations could be in different mental states. Rejecting Twin Earth externalism on the other hand is

---

[8] For a disjunctive conception see for example McDowell 1982 or McDowell 1986.

[9] Again, since internalism comes in many varieties, perhaps not all internalists would agree with putting the doctrine in this way. For an example of the debate conducted in similar terms, see John McDowell's comments on Simon Blackburn in the context of a dispute over object-dependent thoughts. Blackburn describes a series of Twin Earth style scenarios where "everything is the same from the subject's point of view" and claims that "there is a legitimate category of things that are same in these cases" (Blackburn 1984, 324). McDowell agrees as far as "(t)he uncontentiously legitimate category of things that are the same across the different cases is how things seem to the subject" but he denies that there would be "something" - a mental state, for example - which is the same across these situations. (McDowell 1986, 157).

the denial of such possibilities.[10]

The notion I am suggesting is clearly related to the usual understanding. Things in the world normally effect us through effecting the surface of our body. One obvious way to create two subjectively indistinguishable situations for a subject is to keep the proximate stimuli on her bodily surface constant, while varying the causal origin of the stimuli, and in these cases, the external facts are indeed outside the body. So the customarily discussed Twin Earth cases will turn out to be Twin Earth cases according the new interpretation - just as they should.

Moreover, my proposal covers the cases discussed in sections 2 and 3. In the meningitis case and in both versions of the water case, the point of the argument is not that the Twins are molecule for molecule identical (as they are not); the crucial stipulation in the scenario is that the Twins are in subjectively indistinguishable situations.

Externalist critics of the Cartesian theory of the mind often identify the Cartesian description of the evil demon (or the brains in a vat) hypothesis as a central feature of the theory. The claim is that even if you were deceived by an evil demon, or were a brain in a vat, your thoughts nonetheless would be the same - only most of them would be false. The feasibility of the whole hypothesis depends on what I identified as the key relation in setting up Twin Earth scenarios: the relation of subjective indistinguishability; because if you were a brain in a vat, *everything would appear the same*. Externalists are divided over the question of what to say about vat-brains. Putnam grants the intelligibility of the hypothesis; his externalism is manifested in the claim that contrary to the Cartesian assumption, the thoughts of vat-brains would be different from our thoughts. Other externalists question the intelligibility of the whole scenario. In any case, disagreement over the evil demon or the brains in a vat hypothesis between internalists and externalists is a disagreement over what subjective indistinguishability implies, and this makes it immediately clear why the Cartesian theory is an internalist theory.

Understanding the relation between the Twins in terms of subjective indistinguishability is also applicable to other brands of externalist arguments. Putnam's argument from the division of linguistic labor and Burge's argument for social externalism both involve imagining two linguistic communities where the use of certain expressions differ. Then we are to place a Twin in each of these communities, and according to the argument, they will have different concepts. These arguments would *not* obviously be arguments for *externalism* if the Twins somehow *registered* the

[10] It may be objected that the thesis of indistinguishability as the criterion for identity of mental states makes internalism a non-starter, because of the intransitivity of phenomenal indiscriminability. I don't think this issue is settled: for an argument for the transitivity of this relation see Jackson and Pinkerton 1973 and Graff 2001. The main problem I see with the denial of transitivity is, very briefly, this: the relation 'same appearance' has to be transitive (since it's based on the *identity* of appearances). Denying the transitivity of indiscriminability therefore commits one to denying that indiscriminable situations have the same appearance. This is possible but I think undesirable.

relevant differences in usage. The crucial assumption of the scenario is again that if the Twins were counterfactually swapped, *the situation would be indistinguishable for the subject*.

5. Incompatibility and the usual understanding

The significance of this proposal lies in the fact that it alone helps to understand why anyone should have thought that *unlike internalism*, externalism is incompatible with self-knowledge or privileged access. First I shall argue that on the usual understanding of externalism, we can expect no significant difference between externalism and internalism in relation to self-knowledge.

Suppose that we accept the usual understanding of externalism, which draws the boundary between the internal and the external around the skin (or the brain). Then the main difference between the internalist and the externalist is about where to locate facts on which the content of our mental states depend:

externalism:

being in a mental state with content *C*

depends on/entails that

E (some fact which is outside the body or the brain of the subject)

internalism:

being in a mental state with content *C*

depends on/entails that

B (some fact about the body or the brain)

The idea that externalism is incompatible with privileged access is usually articulated with the help of contrasting our epistemic status with respect to the first and the second item in the externalist thesis. Thus we know in some special way (directly or a priori or with first-person authority or something like that) that we are in mental state with content *C*, but we do not know in that special way that *E* obtains. And how could something that we know in that special way depend on or entail something we do not know in that special way? The details of the argument are filled in according to what we take to be the 'special way', and according to what we take to be the nature of 'dependence' or 'entailment' between the first and second item. Witness Burge's formulation of the problem in his influential article defending the compatibility thesis:

> Our problem is that of understanding how we can know some of our mental events in a direct, nonempirical manner, when those events depend for their identities on our relations to the environment. A person need not investigate the environment to know what his thoughts are. A person does have to investigate the environment to know what his environment is like. Does this not indicate that mental events are what they are independently of the environment? (Burge 1988, 650)

But if this is indeed the source of concern about compatibility, then the internalist has as much reason to worry as the externalist has. Consider the formulation of internalism above: the same contrast can be drawn between our epistemic status with respect to the first and second item in the thesis. We certainly do not know directly and non-empirically our brain-states, nor, under a similar description, the bodily states which are meant to individuate our mental states. We find out many things about our body in the same way we find out things about our environment: empirically and from the third-person point of view - with the help of X-rays and surgery and tissue-samples. If the *only and decisive* difference between internalism and externalism is whether they place facts that individuate mental content within or outside the confines of the body, there is no reason to think that this will result in any interesting epistemological difference between the two theses.

Burge makes use of this insight in his criticism of the idea that externalism is incompatible with self-knowledge. On his understanding, the argument for incompatibility has the same root mistake as Descartes' argument for the real distinction between mind and body. Granting that we know our mental states in a special way does not entail that we know *everything* about them in the same special way - it still leaves room for the claim that those states depend on facts about the body or facts about the environment.

> One can know what one's mental events are and yet not know the relevant general facts about the conditions for individuating those events. It is simply not true that the cogito gives us knowledge of the individuation conditions of our thoughts which enables us to "shut off" their individuation conditions from the physical environment. (Burge 1988, 651)

I think that Burge is essentially right on this.[11] It is somewhat puzzling though why so much time was spent on arguing for or against the incompatibility of self-knowledge and externalism, if the solution to the problem is so simple. I suggest the following explanation: the solution is simple only if we rely on the usual understanding of externalism - in that case, externalists and (materialist) internalists have indeed as much or as little reason to worry about compatibility with self-knowledge. But this is not the last word in the debate: for on my new understanding, there *is* a difference between internalism and externalism in their relation to self-knowledge. I think that a tacit

[11] Burge also presents a positive theory of self-knowledge which he claims to be compatible with externalism (for a development of the theory, see also Burge 1996). The essence of the theory is that second order thoughts like 'Now I am thinking that water is wet' are contextually self-verifying: since the content of the second order thought inherits the content of the first-order thought 'water is wet', there is no possibility of mismatch between the two contents. I do not think this solution is satisfactory. I cannot go into details here, but the main problem seems to be that on Burge's theory, the correctness of second order thoughts is a result of their contextual character; essentially in the same way as I cannot be wrong in thinking that I am here. This latter, however, is compatible with my knowing nothing about where I am. It is arguable that in the case of self-knowledge we have this second type of more substantial knowledge - and therefore Burge's theory does not account for the entire scope of self-knowledge.

reliance on something like the conception I suggest could explain the persisting feeling that there is a problem here. Before I show this, let me discuss briefly another popular version of the incompatibility argument.

6. The McKinsey argument

A frequently discussed form of the incompatibility argument follows a pattern first suggested by Michael McKinsey. Here is the gist of the argument:

> ... if you could know a priori that you are in a given mental state, and your being in that state conceptually or logically implies the existence of external objects, then you could know a priori that the external world exists. Since you obviously don't know a priori that the external world exists, you also can't know a priori that you are in the mental state in question. It's that simple. (McKinsey 1991, 16)

The argument is a reductio: the claim that self-knowledge is a priori, combined with the externalist thesis, leads to the unacceptable conclusion that we know a priori certain facts about the external world. What is essential to this argument - but, as we shall see, also highly controversial - is that the externalist thesis and its specific application in the argument should be known a priori in a sufficiently strong sense; otherwise the empirical presuppositions of the thesis may explain the empirical nature of the conclusion. But whether this can be shown or not, the main problem again is that if we accept the usual understanding, internalists seem to have as much or as little reason to worry about the McKinsey argument as externalists do.

To see this, let us apply the reductio to *internalism* as formulated above (according to the usual understanding). On one version, we have the conclusion that we know a priori that our brains exist, which is clearly false. It will be said that this is because in-the-brain internalism is not wholly a priori, being based on the empirical assumption that we have a brain. True enough; then obviously outside-the-brain externalism is not wholly a priori either. This latter theory states that facts outside our brain individuate mental contents - so the theory relies on the empirical assumption that we have a brain.

On the other version, we have the conclusion that we know a priori that a certain bodily state exists. This poses a question which I won't be able to discuss here in proper detail: do we know a priori that our body exists? If the answer is no, then the conclusion of this argument is again as unacceptable as the conclusion drawn from the externalist thesis, and we have a reductio against the compatibility of materialist internalism and privileged access. Now just as before, the empirical content of the conclusion may be the consequence of some empirical assumption in the internalist thesis. However, the same assumption - namely that we have a body and it marks a relevant

boundary in locating facts - will also spoil the a priori character of the externalist thesis.

But perhaps it will be suggested that the internalist thesis won't run into the same difficulties as the externalist claim, because we *do know* a priori that our body exists. What this means is not entirely clear: it sounds odd to say that we know that our body exists independently of or without experience. Therefore we would need some other sense of the 'a priori', and this is where one might start to doubt that appealing to this notion will be useful in this context. In any case, it seems likely that on any interpretation of 'a priori' that makes it plausible that we know a priori that our body exists, it will be arguable that we know a priori that objects outside our body exist. For example, it may be held that the fact of our embodied existence is *part of our conceptual scheme* and hence a condition for any experience; but there is nothing outlandish or absurd about the claim that the same is true of the existence of material objects outside us.

There are many versions of the McKinsey argument, and some of them may avoid the problems mentioned here, but I can't provide a detailed discussion here. Let me just say that I am not convinced that focusing on the putative a priori character of self-knowledge is helpful in this context. One reason for this is that the notion of a priori does not enjoy universal acceptance, since there are a number of philosophers who are convinced by Quine that there is no purely a priori knowledge. However, should Quine be right on this, it seems to me that one could still argue for the special nature of self-knowledge.[12] Second, even if we resist Quine's conclusion, it's still not obvious that self-knowledge is a priori on any plausible understanding of the notion. As far as I can see, the best explanation of the a priori that has been provided so far is in terms of analyticity[13], but that's not applicable to self-knowledge. Instead, we are left with somewhat unspecific terms like 'without empirical investigation of the world' or 'by thinking alone'. These phrases in fact say hardly more than that is *some* difference between the way we know the world and the way we know our thoughts.[14]

7. Externalism and privileged access

If we accept my understanding of externalism, we will have an argument that makes privileged access incompatible with externalism but not with internalism. First we should get clear about the features of privileged access which generate the incompatibility. We have already encountered the suggestion that what makes self-knowledge privileged is its *a priori* character; but for the reasons given above, I do not think this is helpful. There are other customarily held features of privileged

[12] Davidson seems to be a case in point: see Davidson 1987.

[13] As for example in Boghossian 1996.

[14] Another source of problems is the putative a priori character of the externalist thesis. For similar doubts and a convincing argument that no notion of the a priori will serve the purposes of the incompatibility argument, see Nuccetelli 1999.

access, and I recommend to focus on *first-person authority*. Having first-person authority about my thoughts does not necessarily mean infallibility about them; it means only that I am in a better position to know my own thoughts than anyone else.

Privileged access, when characterized in this way, is plausible primarily about occurrent thoughts and experiences. Explaining knowledge of our beliefs, desires or intentions requires a more complicated story: phenomena like self-deception, difficulty of grasping complex ideas or the effects of strong emotional involvement suggest that such states are often not known with first-person authority. For reasons like this hardly anyone would want to maintain that we have unrestricted privileged access to *all* of our mental states. The striking feature of externalism is that it forces a limitation on privileged access which is *fundamentally different* in character: it arises with respect to the simplest occurrent thoughts and experiences, and it is not explainable by these familiar facts of human psychology. This is an important point which is often overlooked by externalists: they simply list examples (like the above) where we have limited self-knowledge, and then effortlessly extend the limitation to cases which are clearly quite different.

The incompatibility of first person authority and externalism (in my understanding) is quite straightforward. Externalists hold that a subject in subjectively indistinguishable situations could have different mental states. But first-person authority extends only as far as things are subjectively distinguishable, that is, distinguishable from the subject's point of view. If I would never notice the difference between this situation and a Twin situation, then of course other people could be in a better position to detect the difference. Internalism, on the other hand, poses no such restrictions on the scope of first-person authority: for on this view, everything which could make a difference to a subject's mental states should be discriminable by the subject herself. The internalist does not have to insist that *every* fact our thoughts depend on - the existence of our brain, for example - can be registered by first-person authority. Nonetheless she does insist that it is legitimate to claim that facts individuate mental contents only insofar as they *make a difference* to the way things appear to us. This means that any difference in the content of thoughts should be distinguishable from the subject's point of view and hence remains within the reach of privileged access.

8. Metaphysics of the mind

I anticipate a certain objection. Someone could say that if I define externalism and privileged access in *this* way, the incompatibility immediately falls out; but then my argument is simply question-begging. I think that in a sense this is right: one way to sum up my proposal is to say that externalism is a thesis about the nature of our access to our thoughts. Yet it is important to see that nothing I said settles the outcome of the internalism/externalism debate. The arguments presented in this paper leave a number of options open. You can choose to ignore the points about dualism and the meningitis case and hold on to the usual understanding of externalism and internalism; but in that

case, you should not expect there to be a special issue about self-knowledge between externalism and internalism. Alternatively, you can accept my understanding, insist on the correctness of externalism, and conclude that we do not have first-person authority over our thoughts. This doesn't necessarily mean giving up the thesis of privileged access altogether; one could still to try to account for the privileged nature of self-knowledge in terms of some other feature. Finally, you could again accept my understanding, and argue that since first-person authority is an essential feature of knowledge of our thoughts, and externalism is incompatible with that, internalism wins.

This may not satisfy those who see my argument as question-begging, so let me conclude with considering an objection of this sort which in fact goes to the very heart of the matter. The objection may go like this: 'One way to explain why the issue of self-knowledge and externalism proved to be so difficult is to point out the different nature of the two doctrines. Externalism and internalism, being theses about content-individuation, are *metaphysical* doctrines and hence should be cast in metaphysical terms, whereas the thesis of privileged access is an *epistemological* doctrine, formulated in epistemic terms. What makes the demonstration of incompatibility hard - or even impossible - is the difficulty of drawing epistemic consequences from a metaphysical distinction. In fact, this was illustrated quite well by what was said about the hopelessness of the incompatibility argument assuming the usual understanding. It is not surprising that *your* demonstration of the incompatibility was so easy: that's because you illicitly defined externalism and internalism in terms of subjective indistinguishability, that is, in *epistemic terms*. But then you failed to draw the relevant metaphysical distinction.'

Keeping apart metaphysical and epistemic issues is usually a good policy, but I don't think that a strict separation is feasible when our interest is the *metaphysics of the mind*. What it is to have a mind is inseparable from what it is for example to have experiences, and this latter is a thoroughly epistemic notion. How metaphysics and epistemology are intertwined in philosophizing about the mind can be illustrated by countless examples from Descartes to Sellars. This is especially true when the question is about knowing our own mind; as Colin McGinn put it, '... we cannot *first* fashion a conception of the mind and *then* go on to specify the ways in which the mind is known. In a word, there is no epistemologically neutral conception of the mind ...' (McGinn 1982, 7).

In some discussions of externalism, it is in fact made clear that opting for externalism or internalism turns on rejecting or accepting some epistemic assumption. A good example is John McDowell's discussion of object-dependent thoughts (in McDowell 1986). A thought is object-dependent if it cannot exist without its object existing. On the original Russellian conception, the class of such thoughts is limited to thoughts about sense-data and to thoughts about ourselves. McDowell suggests to extend the class to include certain thoughts about external physical objects, hence arriving to an externalist conception of the mind. He makes it clear that

> Russell's restriction results, in effect, from refusing to accept that there can be an illusion of understanding an apparently singular sentence... (138)
> If we lift Russell's restriction, we open the possibility that a subject may be in error about the content of his own mind... (145)

What motivates McDowell's externalism is the conviction that it provides the only way to account for the relation between mind and world: that our thoughts are *about* objects in the external world. But he also realizes that adopting externalism immediately involves a restriction on the access to our thoughts. And I think that opposition to externalism may well emerge from refusing to compromise on this question.

References

Blackburn, S. 1984: *Spreading the Word* Oxford: Clarendon Press

Boghossian, P. A. 1996: "Analyticity Reconsidered" *Noûs* 30/3: 360-391

Boghossian, P. A. 1997: "What the Externalist Can Know A Priori" *Proceedings of the Aristotelian Society* XCVII/2: 161-75 (also in Wright-Smith-MacDonald 1998)

Burge, T. 1982: "Other Bodies" in: Andrew Woodfield (ed.): *Thought and Object*. Oxford: Clarendon Press: 97-120

Burge, T. 1986: "Cartesian Error and the Objectivity of Perception" in Pettit-McDowell 1986: 117-36

Burge, T. 1988: "Individualism and Self-Knowledge" *Journal of Philosophy* 85: 649-63.

Burge, T. 1996: "Our Entitlement to Self-Knowledge" *Proceedings of the Aristotelian Society* XCVI: 91-116

Davidson, D. 1987: "Knowing One's Own Mind" *The Proceedings and Addresses of the American Philosophical Association* 60: 441-58 reprinted in Cassam, Quassim (ed.) 1994: *Self-Knowledge*. Oxford University Press: 43-64

Davies, M. 1998: "Externalism, Architecturalism and Epistemic Warrant" in Wright-Smith-MacDonald 1998: 321-61.

Graff, D. 2001: "Phenomenal Continua and the Sorites" Mind 110: 905-935

Jackson, F. and Pinkerton, R.J. 1973: "On an Argument against Sensory Items" *Mind* 82: 269-272

Jackson, F. and Pettit, P. 1996: "Functionalism and Broad Content" in Pessin, A. and Goldberg, S. (eds.) *The Twin Earth Chronicles* M.E. Sharpe: 219-30

McCulloch, G. 1995: *The Mind and its World*. London and New York: Routledge

MacDonald, C. 1998: "Externalism and Authoritative Self-Knowledge" in Wright-Smith-MacDonald 1998: 124-54

McDowell, J. 1982: "Criteria, Defeasibility and Knowledge" in Dancy, Jonathan (ed.) 1988: *Perceptual Knowledge* Oxford University Press: 209-19

McDowell, J. 1986: "Singular Thought and the Extent of Inner Space" in Pettit-McDowell 1986:

136-168
McGinn, C. 1982: *The Character of Mind* Oxford, New York: Oxford University Press
McKinsey, M. 1991: "Anti-Individualism and Privileged Access" *Analysis* 51.1: 9-16
McLaughlin, B. P. & Tye, M. 1998: "Externalism, Twin-Earth and Self-Knowledge" in Wright-Smith-MacDonald 1998: 285-320
Nuccetelli, S. 1999: "What Anti-Individualists Cannot Know A Priori" *Analysis* 59.1: 48-51
Pettit, P. and McDowell, J. 1986: *Subject, Thought and Context* Oxford: Clarendon Press
Pettit, P. 1986: "Broad-Minded Explanation and Psychology" in Pettit-McDowell 1986: 17-58
Putnam, H. 1975: "The Meaning of 'Meaning'" in *Mind, Language and Reality* Cambridge University Press: 215-271
Segal, Gabriel 2000: *A Slim Book about Narrow Content* Cambridge, MA: The MIT Press
Wright, C. Smith, B. C. and MacDonald, C. 1998 (eds.): *Knowing Our Own Minds* Oxford University Press

Department of Philosophy
Central European University, Budapest