Does Twin Earth Rest on a Mistake?

«Does Twin Earth Rest on a Mistake?»

by Katalin Farkas

Croatian Journal of Philosophy
Vol. III, No. 8, 2003

Does Twin Earth Rest on a Mistake?

KATALIN FARKAS
Central European University, Budapest

In this paper I argue against Twin-Earth externalism. The mistake that Twin Earth arguments rest on is the failure to appreciate the force of the following dilemma. Some features of things around us do matter for the purposes of conceptual classification, and others do not. The most plausible way to draw this distinction is to see whether a certain feature enters the cognitive perspective of the experiencing subject in relation to the kind in question or not. If it does, we can trace conceptual differences to internal differences. If it doesn't, we do not have a case of conceptual difference. Neither case supports Twin Earth externalism.

I.

'I have never ceased to be surprised and gratified by the speed with which and the extent to which the view I proposed in that essay became widespread' said Hilary Putnam in the introduction to *The Twin Earth Chronicles*, a collection of responses to his 'The Meaning of 'Meaning", published in 1995, twenty years after the original article.[1] I think surprise is indeed the adequate reaction to the incredible influence of the Twin Earth argument, though probably my reasons are different from those of Putnam. The view that Putnam had proposed has subsequently became known as externalism about content, and has proliferated in many versions. My concern in this paper is the variety which is expressly motivated by Twin Earth arguments, and which therefore I shall call Twin Earth externalism. Twin Earth externalism aims to establish—usually through the analysis of a concrete example—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.[2] The aim of this essay is to show

[1] P. xv. in Pessin and Goldberg [1995].

[2] As far as I can see, this is equivalent to the more usual formulation: the claim that the content of (some of) a subject's mental states depends (partly) on facts outside her. That Twins or Doppelgängers can be used to define externalism is suggested for example in Jackson and Pettit [1988], 220, McLaughlin & Tye [1998], 285 and in Davies [1998], 327.

that Twin Earth arguments and Twin Earth externalism rest on a mistake.[3]

II.

The best known Twin Earth scenario is probably Putnam's; I shall come back to this case at the end of the paper; but first I propose to discuss a different Twin scenario presented by another notorious defender of externalism, Tyler Burge (in Burge [1986a]). I chose Burge's story because a discussion of his argument is especially suitable to illustrate what I think is at stake in the externalism/internalism debate: the question is whether features of things make a difference to the concepts we form about them simply by being present—or only when they make a difference to the cognitive or experiential perspective of cognisers.[4]

Burge's first Twin situation is this: the protagonist—call him Sebastian—correctly perceives some objective entities, for example small shadows on a gently coloured surface.[5] Small shadows look the same as similarly sized cracks, and Sebastian sometimes misperceives a crack as a shadow—what is in fact a crack he takes to be a shadow. But he never, on the basis of his experiences, discriminates cracks from shadows, and furthermore, he has no disposition to distinguish them, for example, by touch. Burge thinks that after some experiences of shadows, in this situation Sebastian acquires the concept *shadow*. Next we are to imagine a counterfactual situation, where owing to peculiar optical laws or effects there are no visible shadows. There are, however, plenty of cracks, which look the same and hence cause the same proximate stimuli as shadows do in the normal situation. Each time Sebastian experiences a shadow in the normal situation, he experiences a crack in the counterfactual one. According to Burge, in the latter case, after some experiences of cracks, Sebastian ends up with the concept *crack* (rather than with the concept *shadow*).

It is stipulated in the example that the internal states of Sebastian in the actual and counterfactual situations are the same. However, something is different in the actual and counterfactual environments: in one case, Sebastian's relevant experiences caused mostly by the presence of shadows and only seldom by cracks, and in the other, there are no shadows, and all the relevant experiences are caused by cracks. And this is why, according to Burge, Sebastian ends up with different concepts. Assuming compositionality, the content of Sebastian's mental states will be different in the two situations. Thus we have a case of Twin Earth exter-

[3] My discussion is aimed at arguments similar to those given by Putnam; this means that I shall discuss Twin Earth arguments based on *general concepts* only.

[4] The cognitive and experiential perspective in the present sense supervenes on internal states. For a detailed argument for an understanding of the debate in these terms, see Farkas [2003].

[5] I am slightly modifying Burge's text: he talks about a subject *S* perceiving objective entities *O*s, and small shadows are possible examples of *O*s. I use a name and the examples for easier readability.

nalism, moreover, a Twin Earth externalism which does not appear to rest on the case of natural kind terms.

The natural way to resist the externalist conclusion is to claim that Sebastian acquires the *same concept* in both the actual and counterfactual situations, a concept which has *both shadows and cracks in its extension*. Notice that it follows from the requirement of the sameness of internal states that Sebastian gives the *same name* to the concept in both situations. Let this name be 'crackdow'. Then the internalist opponent of the externalist insists that 'crackdow' expresses the same concept in both situations, the term will be correctly applied to both shadows and cracks, and hence the content of 'crackdow' thoughts will be the same.[6]

What supports Burge's analysis of the example? Unlike with some other versions of the Twin Earth argument, Burge does not mean to present the shadow-crack case to provide an overwhelming *intuitive* support for his externalist thesis. He rather uses the example to illustrate a certain conclusion he draws from what he takes to be plausible premises about perception and perceptual concepts. The premises are as follows (Burge [1986a], 125ff):

(1) our perceptual experience is about objective (i.e. mind-independent) objects, properties or relations;

(2) we have perceptual representations which specify objects etc. as such (so the fundamental analysis of perceptual content is *not* 'whatever causes this sort of perceptual representations');

(3) 'some of a perceiver's perceptual types take on their representational characters partly because their instances interact in certain ways with the objective entities that are represented' (Burge [1986a], 130).

First let me clear a possible misunderstanding out of the way. The last premise can be understood in a way which is very plausible, but doesn't support Twin Earth externalism. In some sense, what we find in our environment will obviously explain what concepts we are likely to acquire. For example, if I am right in thinking that people hadn't had the concept *platypus* before they went to Australia, then we can easily explain this fact by pointing out that they encountered platypuses only in Australia. If this is the whole meaning of the premise—a historical-genealogical account of certain aspects of concept-acquisition—there would be little reason to object to it. The contribution of the external world, however, is considered in a different manner in the externalism-internalism debate. Both sides may agree that some of our concepts are *actually* acquired (partly) through causal interaction with the world; the disagreement is over the question of whether it would have been *possible* to acquire the same concepts without the contribution of the world. An internalist would insist that, even if though it is not likely, it is still possible to acquire the concept *platypus* even if one has no causal interaction with a platypus, or indeed, even if platypuses do not exist.

[6] The term 'crackdow' is from Gabriel Segal's paper (Segal [1989]). Segal convincingly criticises another use of the example by Burge, where Burge tries to show the externalist character of Marr's theory of vision (in Burge [1986b]).

Burge's thought is that since the *objective entities* are *shadows* in one case and *cracks* in the other, this will result in a difference in the representational character of the subject's respective perceptual types. However, I shall argue that in using his premise (3) in this way he actually violates his own premise (2).

III.

One feature of the Twin scenario, mentioned only in passing by Burge, but actually very important, is that Sebastian does not distinguish, and has no disposition to distinguish shadows and cracks by touch or any other means. It would clearly spoil the case if Sebastian, each time when he saw a shadow or a crack, would touch it to find out whether it was a shadow or a crack. (For example, imagine that Sebastian was examining a wall in a lamp-lit room to find out whether the painting of the wall was even.) In that case, proximal stimuli, 'narrow' experiences and hence internal states would be different in the two cases, and consequently the internalist would be happy to agree that Sebastian acquired different concepts in the two situations. Moreover, I think Burge is right: it is not enough if Sebastian does not actually touch the surface; he also has to lack a disposition to do so. We have to assume that the difference between shadows and cracks *does not matter for Sebastian*. The description of the scenario evidently suggests that the concept in question is associated only with a certain visual impression, and there are no further features of the 'objective entities' encountered which matter as to their being instances of crackdows.

Our interest in the world is different from Sebastian's: *we*, unlike Sebastian, do have dispositions to distinguish between shadows and cracks. For something being a shadow or a crack, it does matter what tactile impression it would give rise to. But it is hard to see how, without such dispositions and interests, we would actually have the distinction between shadows and cracks. Therefore it seems rather implausible that Sebastian, having no such interests and dispositions, should end up in either of the two situations with anything like *our* concept of shadow or crack. But then there seems to be no reason to assume that he would acquire different concepts in the two situations.

The difference between Sebastian and *us* is that we make certain distinctions that Sebastian does not make and is not disposed to make. To see that this has no bearing on *Sebastian's concepts*, consider a converse case: someone who discriminates *more finely* than we do. Suppose that the perceived entities in both cases are cracks, the difference between them being that in the first situation, the usually perceived cracks are not deeper than a half centimetre, whereas in the counterfactual situation, all perceived cracks have a depth between half and one centimetres. Let's call these shallow-cracks and deep-cracks. First, consider people like us: we cannot reliably distinguish between shallow and deep cracks by unaided vision or touch, but we don't care about distinguishing them either. Let us stipulate that in one possible history of acquiring the concept crack, we encounter mainly shallow-cracks, and in another possible history, we encounter only deep-cracks. It is hardly plausible that we acquire *two different concepts*: one having only shallow-cracks, the other only deep-cracks in its extension.

From this point of view, it does not even matter if, due to some peculiar effects, there are no shallow-cracks in the second situation. For if there were, they would be still classified as cracks by us. What if the existence of some entities were forbidden by some laws (as the description of Burge's case may suggest?) Discussing this aspect of Burge's scenario—the absence of shadows due to peculiar optical laws—would lead too far. To see clearly on this point, we should know if the laws of nature are in fact different in the counterfactual situation, and if they are, to what extent, etc. Let me just venture one general remark: perhaps if you change the laws to a sufficient extent (provided that you can, i.e., if they are not necessary), this may result in a situation when even an internalist would be reluctant to say that a creature has the same concepts as we do. But it is doubtful whether in such a scenario we can apply the condition of 'internal sameness'. Our internal physical constitution depends on the laws of nature, and if these are very different, the debate on whether internal sameness implies sameness of content becomes pre-empted.

Back to shallow-cracks and deep-cracks: we are not interested in this distinction, but someone *else* might be (for example for the purposes of a quality-check on the painting of a wall: shallow-cracks are allowed, but deep-cracks mean repainting). This means that *she* would be disposed to distinguish between the two kinds. Then it would be right to describe *her* situation by saying that the 'objective entities' she encounters are shallow-cracks in one case, and deep-cracks in the other case.

The mere possibility of an individual with such interests—interests unlike ours—should have no consequences to our case. This is especially true because there could be any number of accidental differences between the 'objective entities' present in the actual and the counterfactual situations, respectively. (There could be old cracks and young cracks, cold cracks and warm cracks, and so on.) And if we took all of these seriously, we would end up with a wildly implausible picture of concept-acquisition and concept individuation: that depending on some accidental features of my actual history, I could—unknowingly, of course—end up with a concept which has only young cold cracks, rather than cracks in its extension.

Interestingly enough, Burge's conclusion appears to go against something else he says in the same paper. When discussing 'Cartesian' thought experiments—about the possibility of our being deceived by a demon or being brains in vats—he warns us about a certain problematic feature of considering counterfactual situations. Naturally, we describe counterfactual situations using our language and our concepts. But when the question is just about what thoughts or concept someone in a counterfactual situation might have, we need extra caution. In Burge's argument, this is directed against the Cartesian: for from the fact that we can imagine a counterfactual situation which would make *our thoughts* radically mistaken (say the brain-in-a-vat scenario) it doesn't follow, according to Burge, that if we *were* in that situation, we would still have the same thoughts. But it seems that in the shadow/crack case Burge forgets his own warning about the dangers of using our own language to characterise a counterfactual situation. He describes the 'objective entities' which contribute to the representational character of Sebastian's concept by using *our* concepts of crack and shadow. Just as someone else might describe the 'objective entities' *we* encounter as shallow-cracks and deep-cracks. No consequences should follow from this concerning either Sebastian's or our own concepts.

Still more interestingly, Burge could have avoided getting into trouble on this score if he had only took his own premises seriously. For Burge uses his third premise in a way that violates his second premise. The second premise says that we have perceptual representations which specify objects etc. as such. But the 'as such' requirement makes sense only if we apply it *from the point of view* of the concept-acquirer. The shadow-crack distinction does not apply at all from Sebastian's point of view, so Sebastian's experiences do not specify shadows and cracks 'as such', but simply *as crackdows* (which could be either shadows or cracks).[7] The only way Burge can get his externalist conclusion is if, contrary to (2), he claims that *whatever happens to be there to cause Sebastian's experiences*—shadows in one case, cracks in the other—will determine content.

I shall call this latter idea the 'whatever happens to be there' principle. The principle need not be asserted in its crudest form: that for example hallucinations regularly caused by drugs would represent drugs which caused them; the idea is only that some external element (beside the other possible content-constituting elements) contributes to content not because it figures in the cognitive or experiential perspective of a cogniser, but *simply because it happens to be there*. To put it more formally, if less intuitively, the 'whatever happens to be there' principle says this: if a subject *S* acquires a concept (partly) through interacting with entities in her environment, and most (or all) of these instances *happen* to possess a property F, then the concept she acquires has *only* (though of course not necessarily all) Fs in its extension—*even if* the F-ness of something does not make a difference to *S*'s experiences and she either does not, or cannot, or simply would not discriminate between instances of the concept which are Fs and those which are non-Fs.

This principle is problematic, because of the reasons mentioned two paragraphs back: there are *too many* properties which all the actual instances could, as a matter of accident, share, and which are external (in the above sense) to the subject's cognitive or experiential perspective. And, provided that she doesn't care about the presence of these proper-

[7] It is important to note that on the internalist conception, crackdow is *not* a disjunctive concept which *means* 'shadows-or-cracks'. A concept is not disjunctive merely because its extension falls into exclusive subclasses: the concept 'horse' is not disjunctive merely because horses are stallions, mares or geldings; and the concept crack is not disjunctive because cracks are shallower or deeper than half centimetre.

ties, it is extremely implausible that *all* of them should effect the concept she acquires.

If the principle were accepted, it would be easy to create Twin Earth arguments. We could simply stipulate parallel histories of interacting with entities in our environment, where these entities happen to share or not share some accidental property. But the principle should not be accepted. I think the danger that threatens Twin Earth externalists is a tacit reliance on the principle, as we saw it happening in Burge's case. On the other hand, rejecting the principle does not leave enough room for externalism—as I shall try to show in the following sections.

IV.

Many externalists would probably reply to the concerns raised in the previous section that it is simply not true that they subscribe to the 'whatever happens to be there' principle. Such a claim might be attributed to a defender of a 'crude causal theory', they could say; the theory that whatever causes our representations is going to be represented by them. But no decent externalist defends the crude causal theory.

This latter statement is of course true. In fact, externalist theories depart from the crude idea at least in two different ways. There are those who think that causal connections—or other features of the physical environment—constitute but one factor in individuating content, besides others, which are not reducible to causal or even any naturalistically describable factors. Both Burge and Putnam belong to this set. And there are those who regard the crude causal theory at best as a starting point for naturalising content, and have invested a lot of effort in trying to overcome problems—not unlike the ones mentioned above—which are created by the crude theory.[8] Let us first look at the relationship between the naturalising project and Twin Earth externalism.

In the introduction to a book collecting critical essays on Fodor's work, the editors, Barry Loewer and Georges Rey, offer the following—partly historical, partly problem-oriented—sketch as a background to Fodor's theory of content.[9] First, there were internalist theories. Then Putnam (and Kripke and Burge) convinced many philosophers that internalism was wrong, so the quest for a viable externalist theory started. An early effort was to make *actual causal chains* constitutive of meaning (or content). Consider Putnam's familiar Twin Earth story: we have Oscar$_1$ living on Earth, and his molecule-by-molecule replica, Oscar$_2$, living on Twin Earth. The only difference between Earth and Twin Earth is the chemical composition of the liquid known as 'water' in the two planets; it is H_2O on Earth and XYZ on Twin Earth. The idea behind the actual-historical account is roughly this: some original dubbing started a causal chain which led to my present use of 'water'; and since it was H_20 and not

[8] See for example Jerry Fodor's [1987] *Psychosemantics* and 'The theory of content I, II' in Fodor [1990]. The term 'crude causal theory' is from chapter 4. of the former.

[9] Pp. xxiv-vv in Loewer and Rey [1991].

XYZ which was at the end of the causal chain, I have the concept water rather than twater. However, the actual-historical account creates certain difficulties: what determines, for example, that when I dub an animal as ‘duck’, it is *ducks* and not *birds* or *animals* or *that particular duck* that got dubbed?[10]

Loewer and Rey offer the following solution: ‘… a natural answer to the question of what a dubber dubbed might be: whatever kind of thing she would *discriminate* as that thing; that is, whatever she would apply the thing to, as opposed to everything she wouldn’t. Along these lines, a number of philosophers have considered ways in which states involving token expressions might have, in addition to *actual* causal histories, certain *counterfactual* dispositional properties to covary with certain events or properties in the world.’ (ibid., p. xxv)

Fodor’s view in ‘The theory of content’ (which is the main target of several critical contributions to the collection) is a specific version of this mixed view. It is ‘mixed’, because actual causal histories as well as counterfactual covariation individuate content. As long as actual histories are (at least partly) constitutive of concepts, there is a difference between causal chains which start with XYZ or H_2O, and one is likely to fall victim to Twin Earth externalism.

However, Fodor apparently had second (or considering his earlier work, third or fourth) thoughts about this. In an appendix to his [1994] *The Elm and the Expert*, he takes up the question again. According to the informational semantics he continues to defend in the book,

> content depends on nomic relations among properties; to a zeroth approximation, ‘water’ means water in my mouth because *being water* and *being disposed to cause my ‘water’ tokenings* are nomically connected properties of water. Presumably such nomic connections could be in place even if none of my ‘water’ tokens have ever *actually* had water as its cause. Presumably, indeed, they could be in place even if there had never been any ‘water’ tokens, or any water. Where nomic relations are the issue, *actual* histories drop out and what counts is only the counterfactuals. (p. 115)

Given the option of leaving out actual histories, Fodor declares that the mixed view is unaesthetic, it is unattractive for further reasons, and argues against historical determinants in the metaphysics of content. ‘Bother Twins!’[11] Whether this means that Fodor himself finally turns his back on Twin Earth externalism (as defined above) is unfortunately not entirely clear if we consider other things he says in the book.[12] Anyhow, I

[10] This is sometimes called the *qua* problem; see also Devitt and Sterelny [1987], sect. 4.4. The problem is a version of problems created by the ‘whatever happens to be there’ principle.

[11] Op.cit. p.116

[12] For some reason Fodor thinks that even on the pure nomic account, me and my Twin should have different concepts expressed by water tokenings. Since he (rightly) considers this as being at odds with the rest of his theory, he decides to undercut Twin Earth externalism by claiming that XYZ is not nomologically possible.

shall now leave Fodor and try to draw a more general moral from what has been said so far.

The objection to the actual-historical account mentioned above clearly has the same root as my objection had against the 'whatever happens to be there principle'. We might dub ducks as 'ducks', but since ducks also happen to be birds, or male or female ducks, or a specific set of ducks, what determines the property which is relevant to the extension of the concept? The idea that seemed to promise a solution to the dubbing problem was a reference to what the subject would discriminate as a certain kind of thing. This is very much in the spirit of our discussion of the 'crackdow' case. The main argument for treating Sebastian's concept as including both shadows and cracks in both situations is that Sebastian is not prepared to discriminate between the two kinds: taking into account Sebastian's dispositions to classify things appeared to be a much better guide to his concept than taking into account actual causal histories.[13]

It seems that in either case the best way to overcome problems raised by actual causal histories is to turn to counterfactuals. But once we have counterfactuals, we might as well drop actual causal histories altogether, that is, give up the 'whatever happens to be there' principle. Assume for a moment that XYZ is nomically possible. In a Twin case, because internal sameness is required, instances of XYZ *would* cause tokenings of the same type of internal representations as instances of H_2O. And conversely: even if all my 'water' tokenings are actually caused by XYZ, it is still true that instances of H_2O have the property of being disposed to cause the same type of tokenings. If content depends on nomic connections rather than on the actual environment I causally interact with, then as long as my internally same Twin is living in a world with the same nomic connections, we will share contents in spite of our (possibly) different environment. Maybe there is some sort of externalism still present in this theory: in the sense that the content of representations depends on nomic connections in the world. This is, however, not Twin Earth externalism. There is, of course, the possibility I indicated earlier: that laws in my Twin's world are different. But it's not clear how, in such a world, I could have a Twin with the same internal states.[14] There seem to be two morals from this case: one is that by giving up the 'whatever happens to be there' principle externalism fades away; the second is that informational semantics is not necessarily Twin Earth externalist.

[13] Appealing to discriminating dispositions doesn't imply an attempt to *reduce* concepts entirely to dispositions. That is, we need not commit ourselves to the claim that all and only those things which the subject would discriminate as belonging to the concept do in fact belong to the extension of the concept. One obvious objection against that view is that subject might be wrong. We could appeal to dispositions without trying to reduce concept possession to them, just like non-naturalistically minded externalists appeal to causal relations without trying to reduce concept possession to them.

[14] Ignoring a possibility once mentioned by David Braddon-Mitchell: that my Twin lives in a 'nomic cocoon', where inside the cocoon the laws are the same as here, and thus allow internal sameness, but the laws outside are very different. I'm not entirely sure I can get my mind around this possibility, but in any case, I'm quite happy if I can defend internalism for all cases except for nomic cocoons.

V.

We have seen two kinds of cases so far: in the first, Burgean story, Twin Earth externalism was defended, by paying the price of subscribing to the implausible 'whatever happens to be there' principle; in the second case, in the hypothetical development of informational semantics, rejection of the 'whatever happens to be there' principle led to a disappearance of Twin Earth externalism. The last issue to be considered is whether Putnam's original argument is more successful in steering a middle course: preserve Twin Earth externalism without a reliance on the 'whatever happens to be there' principle.

Putnam's 'water' case interestingly differs from Burge's case at least in two respects. First, many philosophers apparently think that it is simply *intuitively obvious* that our word 'water' refers exclusively to H_2O, while the Twin Earthian term refers exclusively to XYZ. And this basic intuition poses a serious challenge for any attempts to give an internalist analysis of Twin Earth scenarios. Sometimes it feels as if some philosophers would be quite happy to confess up to internalism—if it wasn't for the vexatious problem of the intuitive force of the water case.[15] I find this reference to intuitive support somewhat baffling, given the number of people I have met who didn't appear to share this intuition at all.

However, I don't think much turns on this. For as we know, the claim about reference is also supported by a theory of natural kind terms, defended in the works of Kripke and Putnam. And here lies perhaps a second difference between the shadow-crack case and the water case. An important element in my argument above was that the shadow/crack distinction is on the same footing as the shallow/deep crack distinction or the warm/cold crack distinction. That is, there is no more initial reason—without considering the subject's cognitive or experiential perspective—to think that the first distinction contributes to the individuation of our concept than there is reason to think that the last two do the same. However, the water case might tempt us to think differently on this issue: the distinction between H_2O and XYZ seems to have more intrinsic metaphysical weight than the distinction between say cold and warm water. And this of course makes a difference to my argument against Twin Earth externalism, which was based on the idea that once you allow *one* accidentally shared property to individuate your concepts, you have no right to refuse to allow *all of them*.

The obvious reply to this second objection is that even 'natural' properties exemplified by members of a natural kind may be too numerous. We had the example of a duck which is also a bird and which, we may add, could also be a ring-necked duck or a tufted duck (apparently, these look

[15] See for example Paul Boghossian: '*widespread intuitions* seems to have it that whereas Oscar's utterance of "Water is wet" expresses a thought that is true if and only if H_2O is wet ..' (Boghossian [1994], 34) In another paper he says that philosophers embrace externalism 'because their intuitive responses to a certain kind of thought-experiment—Putnamian Twin-Earth fantasies—appear to leave them little choice.' (Boghossian [1997], 163)

similar). Or recall some of the replies to Putnam's argument pointing out varying chemical composition even in what we normally take to be water (heavy water and suchlike). This means that even for natural-kind concepts, we could design Twin Earth cases with counter-intuitive results. If one Twin encounters mostly ring-necked ducks, the other tufted ducks, we should attribute different concepts to them, and not the same *duck* concept, even if they don't care about this difference and they are not disposed to discriminate between the two kinds.

There may be a last resort for the externalist. He could say that in the case of natural kind concepts, the *infima* or *lowest species* provides the right conceptual boundaries. And since you cannot go lower than the lowest species, we can get rid of the further proliferation of implausible concepts. But this move won't help. For on this picture, it would be possible to form natural kind concepts only of the lowest species, and this is clearly not true. Ducks form a natural kind, and we can have the concept *duck*, even if this is not a lowest species.

VI.

Let us see what is exactly happening in the 'water' case. H_2O and XYZ share what Putnam calls the stereotype of water: 'a colourless, odourless, tasteless liquid, which quenches thirst, flows in the rivers etc.' If falling under the stereotype—as opposed to possessing a certain underlying structure—were the only criterion for being water, then H_2O and XYZ would be covered by the same concept. For suppose that the Twins regard underlying structure as unimportant. In that case, the fact that instances of this liquid happen to be H_2O in one case and XYZ in the other should not be relevant in shaping the Twins' concept. To think otherwise would involve committing the same mistake Burge made in the shadow/crack case.

This means that the Twins must regard underlying structure as relevant to waterhood; that is, neither Twin would be prepared to regard something as water if it had a different underlying nature than the stuff he has in his environment. This is another way of putting a point often made in the discussion: the argument works only if we attribute referring intentions to the Twins which are specifically associated with natural kind terms.[16] The referring intention would be based on a certain view of natural kinds which Putnam sums up as follows: 'natural kinds are classes whose normal distinguishing characteristics are 'held together' or even explained by deep-lying mechanisms'.[17] The crucial point is that the Twins must intend the term 'water' to refer to a natural kind in this sense,

[16] See for example Boghossian again: one of the presuppositions of the Twin Earth thought experiments is that 'the word "water"—whether on Earth or Twin Earth—must be thought of as aiming to express a natural kind concept' (Boghossian [1997], 165) Similarly, Greg McCulloch points out the importance of a (possible or actual) community 'who understands substance-words in the way Putnam describes' (rather than deciding to use 'water' to apply to everything which superficially resembles water). (McCulloch [1995], 173)

[17] Putnam [1970], 139.

otherwise sameness of underlying structure will not be relevant to the sameness of kind. This is consonant with the internalist position: we can expect features of a kind to effect our concepts only insofar the presence of these features, in relation to the kind in question, enters the cognitive perspective of the experiencing subject.

Many philosophers seem to agree that *we* for instance have the kind of referring intention associated with the term ‘water’ which gives relevance to underlying structure. So far so good, but our case is not really useful for the externalist argument. We have some further ideas about what that underlying nature might be, and it would be difficult to isolate our referring intention from this fairly specific conception of the underlying nature: i.e. *that water is composed of* H_2O. This connection would at least explain the deeply rooted intuition (if there is one) about there being no water on Twin Earth. But since precisely this is supposed to be different on Twin Earth, we are not anymore internally identical to our Twins. If we think that there is no water on Twin Earth because there is no H_2O, this is surely compatible with internalism.

Hence the suggestion in Putnam’s paper to go back to the time when the chemical composition of water was not known. Suppose that people used the term ‘water’ with the intention of referring to the liquid which has the same underlying structure as the liquid in their environment (or they might say ‘the same structure as *this* liquid’, pointing to a glass of water), but they didn’t know what that underlying nature was. This is a point Putnam expresses by saying that there is a hidden indexical component in ‘water’.

If people had absolutely no idea about what the underlying structure was, it is doubtful that such intentions result in anything like determinate reference. In the case of employing an ordinary indexical expression, say ‘she’, we know what *kind* of thing the referent of our expression is—a person or perhaps an animal -, and so it is clear when we should count two uses of the expression as having different referents. Even when we use a more neutral expression like ‘this’, the referring intention is usually expressible by adding a sortal concept: ‘this table’ or ‘this ship’. But when we are left entirely in the dark about what the ‘structure’ might be, it is difficult to conceive how the referring intention could, *in advance*—that is, before the actual identification of the underlying composition—legislate about relevant and irrelevant differences in structure.

Therefore we should characterise the relevant referring intentions as follows: people have a fairly clear idea about molecular structure in general; they regard it as decisive in determining the boundaries of natural substances; they have not identified the specific structure of water yet, but have the intention to refer, by using the term ‘water’, to whatever that shared the molecular structure of *this* liquid. This, I believe, is the best case for Twin Earth externalism. The considerations leading to this case offer very much the same moral as Burge’s case: features like the boundaries of natural kinds or differences in underlying structure are allowed to shape our concepts only insofar the absence or presence of these features is *regarded by concept-users as relevant* to belonging to the extension of the concept. In other words, external features are important only if they are incorporated into the internal cognitive or experiential perspective of cognizers.

There is only one significant difference between this case and Burge's case: the indexical element. In the crack/shadow case 'crackdow'—the concept shared by Sebastian and his counterpart—had the same extension in both the actual and the counterfactual situation. But if 'water' refers to the liquid that has the same structure as *this* liquid, then even an internalist has to admit that the term 'water' has different extension on Earth and Twin Earth—just as it has to be admitted that the extension of the term 'I' is different for Oscar_1 and Oscar_2. After removing all the confusing layers of the story, *Twin Earth externalism boils down to the familiar observation that indexical expressions shift their reference depending on the context*.[18]

VII.

The crucial question is then whether the phenomenon of context-dependent reference is sufficient to support Twin Earth externalism. A widely accepted argument for an affirmative answer would go like this: mental contents are individuated in terms of their truth-conditions. 'I am hungry' thought by Oscar_1 and Oscar_2, respectively, have different truth-conditions—one is true when Oscar_1 is hungry, the other is true when Oscar_2 is hungry. Similarly, 'Water is transparent' have different truth-conditions when thought by Oscar_1 and Oscar_2: truth requires the transparency of H_2O in the first, the transparency of XYZ in the second case. Internally identical subjects thus can have different mental contents.

Does this vindicate Twin Earth externalism? Not necessarily. One could agree that the content of the Twins' thoughts are different, but deny that this is *due to some difference in their external environment*. The truth-conditions for thoughts expressed by sentences like 'I am hungry' differ for Oscar_1 and Oscar_2 because 'I' has a different reference in the two cases. But it would be odd to say that the different reference is a result of something *external* to them. Consider Oscar_1 alone. No matter what changes we imagine in his environment (having water or twater around him, being in love with Lucinda_1 or Lucinda_2, having or not having a duplicate on Twin Earth), his 'I'-thoughts will refer to him, Oscar_1. That Oscar_1 and Oscar_2 refer to different individuals by the term 'I' is simply the consequence of the fact that they *are* different individuals, that is, each of them is the individual who he is. But this is hardly something external, and so

[18] A similar conclusion, though in a different route, is reached in Fodor [1987], chapter 2. (see especially p. 46). Fodor's argument is based on the idea that categorisation for the purposes of scientific psychological explanation is based on a taxonomy of causal powers, and that the relevant causal powers of the Twins' internal states will be the same. The only point left in the Twin Earth argument is the point about indexicals.

it would be rather disingenuous to call this theory ‘externalism’. Moreover, it is possible to develop an internalist account of indexicals, one which explains other indexicals with the help of things bearing a certain relation to the subject. Such theory recognises self-reference as underlying all cases of indexicality, and self-reference, as I just argued, does not support externalism.[19] Therefore the phenomenon of context-dependent reference is not sufficient to support externalism.

VIII.

The mistake on which Twin Earth arguments rest is the failure to appreciate the force of the following dilemma. Some features of objective entities do matter for the purposes of conceptual classification, and others do not. The most plausible way to draw this distinction is to see whether a certain feature enters the cognitive perspective of the experiencing subject in relation to the kind in question or not. If it does, we can trace conceptual differences to internal differences. If it does not, we do not have a case of conceptual difference. Neither case supports Twin Earth externalism.[20]

Bibliography

Boghossian, Paul A. [1994], ‘The Transparency of Mental Content’, *Philosophical Perspectives*, 8, 33-50.

_______ [1997], ‘What the Externalist Can Know A Priori’, *Proceedings of the Aristotelian Society*, 97, 161-75.

Burge, Tyler [1986a], ‘Cartesian Error and the Objectivity of Perception’ in Pettit, Philip and John McDowell [1986], *Subject, Thought and Context* (Oxford: Clarendon Press), 117-36.

Burge, Tyler [1986b], ‘Individualism and Psychology’, *Philosophical Review*, 95, 3-45.

Davies, Martin [1998], ‘Externalism, Architecturalism and Epistemic Warrant’ in Wright-Smith-MacDonald [1998], 321-61.

Devitt, Michael & Kim Sterelny [1987], *Language and Reality* (Cambridge, Mass.: MIT Press).

Farkas, Katalin [2003], ‘What is Externalism?’, *Philosophical Studies*, 112, 187-208

Fodor, Jerry [1987], *Psychosemantics* (Cambridge, Mass.: MIT Press).

_______ [1990], *A Theory Of Content And Other Essays* (Cambridge, Mass.: MIT Press).

[19] Something like this view is defended by John Searle, see Searle [1983], especially chapter 8. For a development of an internalist theory of indexicals along these lines, see for example Ludwig [1996]. An even more radical internalist theory would claim that mental contents can have genuinely context-dependent truth-conditions. Unfortunately, I have no space to work out this solution here.

[20] A version of this paper was presented at the University of Sheffield, and I am grateful for the audience for their comments. I would also like to thank Tim Crane for many discussions on this subject. I am grateful for the OTKA foundation, grant no. T 032435, and for the Hungarian Academy of Sciences, Philosophy of Language Research Group, for support.

_______ [1994], *The Elm and the Expert* (Cambridge, Mass.: MIT Press).

Jackson, Frank & Philip Pettit [1988], 'Functionalism and Broad Content' in Pessin & Goldberg [1996], 219-37.

Loewer, Barry & Georges Rey [1991] (eds.), *Meaning in Mind: Fodor and his Critics* (Oxford UK & Cambridge, Mass.: Blackwell).

Ludwig, Kirk A. [1996], 'Singular Thoughts and the Cartesian Theory of Mind', *Noûs*, 30, 434-60.

McCulloch, Gregory [1995], *The Mind and its World* (London and New York: Routledge).

McLaughlin, Brian P. & Michael Tye [1998], 'Externalism, Twin-Earth and Self-Knowledge' in Wright, Smith, MacDonald [1998], 285-320.

Pessin, Andrew & Sanford Goldberg [1996] (eds.), *The Twin Earth Chronicles* (M S. Sharpe).

Putnam, Hilary [1970], 'Is Semantics Possible?' in *Mind, Language and Reality* (Cambridge: Cambridge University Press, 1975), 139-52.

_______ [1975], 'The Meaning of 'Meaning'' in *Mind, Language and Reality* (Cambridge: Cambridge University Press), 215-271.

Searle, John R. [1983], *Intentionality* (Cambridge: Cambridge University Press).

Segal, Gabriel [1989], 'Seeing What is not There', *The Philosophical Review*, 189-214.

Wright, C., B. C. Smith and C. MacDonald [1998] (eds.), *Knowing Our Own Minds* (Oxford: Oxford University Press).