Surface Externalism
Facing the Dark Side of Twin Earth

David Bourget
February 15, 2005

Putnam concludes his discussion of Twin Earth by proposing his account of grounding\(^1\) for natural kind terms:

The extensions of natural kind terms are determined by ostensive definitions of the form: “‘T’ applies to whatever has the structure producing these superficial characteristics.”\(^2\)

This account is motivated almost exclusively by its capacity to explain our judgment about XYZ. Since \(\text{H}_2\text{O}\) and XYZ do not differ in their superficial characteristics, Twin Earth seems to show that natural kind terms are indexical and that natural kinds are individuated by their causally relevant deep structure.\(^3\) These two claims constitute what I will call the Minimal Putnamian Account (MPA). There appears to be a general consensus that MPA must be true if our judgment about XYZ stems from overarching principles. It thus seems that an independent refutation of MPA and the many proposals that subsume it would suggest that whatever judgments we make about such contrived cases as Twin Earth should be taken as noise and not data to be explained.\(^4\) Sharing in this impression, some have tried to undermine the externalist thesis by criticizing MPA on the basis of imaginative thought experiments that suggest that natural kinds are not individuated by their deep structure (for example, Bach [1987], Mellor [1987]).

---

\(^1\) By “grounding” I mean initial reference-fixing, as opposed to fixing the extensions of one’s terms by borrowing them from others. See Devitt & Sterelny [1999] for a discussion of this distinction.

\(^2\) This reading of Putnam makes him a causal descriptivist. I have no doubt that this is the correct reading of “The Meaning of ‘Meaning’”, but what is even more clear is that this is the only reading worth discussing. A pure causal theory is untenable in light of the qua problem. See Sterelny (1996). The considerations adduced in the present paper also suggest that reference-fixing must be to a great extent “internalized” (albeit not entirely).

\(^3\) The structural properties relevant to MPA are an object’s structural properties which cause or produce its appearance. Not all the structure inside it is relevant. But since this is the standard interpretation of claims about deep structure in this context, I usually drop the qualification to avoid clutter.

\(^4\) To my knowledge, all the causal and causal-historical semantic theories available today subsume MPA.
[1991], and Segal [2000, ch. 5]). My aim here is to respond to this underestimated criticism. I expect that we will come to appreciate important facts about externalism in the process. I first discuss two thought experiments that best bring out the issue.

Suppose that we discover that the entire world is in fact a big text: with a sufficiently powerful microscope, one can see the characters which form all physical objects and substances. For example, it turns out that the watery stuff on Earth is composed of the sequence E-A-U repeated innumerable times in semi-transparent blue letters. Suppose also that it turns out that every white dot in the night sky is an almost perfect replica of our planet. These planets differ from Earth only at the level of deep text. On one planet the sequence of semi-transparent blue letters composing the watery stuff is W-A-S-S-E-R, on another it is W-A-T-E-R, and so on. The same pattern holds for all natural kinds. Further research into the great blueprint of the world reveals that these terms reflect the native languages of the creators, who ended up repeating exactly the same “story” everywhere, albeit using different physical symbols. Should we call W-A-T-E-R “water”? Let me emphasize that the stories of all planets are identical throughout the universe. Mixes are even possible. So Earth speakers can go to Twin Earth 5 and drink A-G-U-A: it takes the place of E-A-U in their bodies without a glitch. MPA predicts that W-A-T-E-R is not water, but this is certainly not what we would say.

Next story. Since cats are not robots on Earth, Putnam suggests we would not classify robot cats discovered on another planet as cats. This would also be my opinion if these robots looked like the latest Japanese toys. But we can imagine other kinds of robots. Say, for example, that robot cats differ from ours only in that they do not have DNA. Instead, they have swarms of picobots which form double-helix structures and realize all the (relevant) functional properties of DNA. Robot cats and normal cats are indistinguishable above the molecular scale. They can even interbreed (engendering normal ones). What is more, the picobots in question are at the origin of life on Earth: they have the capability to create simple molecules which do their work in their place, continuing the proliferation of

---

5 This aspect of the story comes from Bach (1987, ch. 13).
6 I should also specify that other natural kinds found around here have their substance writ in semi-transparent blue letters. This preempts the objection that the property of being composed of blue letters produces wateriness on the present scenario.
species they are programmed to create from scratch. It seems to me that the robot cats I just described are cats. One must take the time to imagine oneself in the company of one of them.

Fully convincing or not, the preceding thought experiments throw a shadow on Putnam’s claim that natural kinds are individuated by their deep structure. If MPA is false and there is no other explanation of the judgment that XYZ is not water, it must be noise in the data. There are many ways our wetware may yield judgments that do not reflect its ruling principles, so externalists need clearer cases to support their sweeping claim.

Now that we glimpsed at the dark side of Twin Earth, let us turn to the bright side. There are two things externalists should try at this point. First, they should try to produce a viable account of grounding which does not make us systematically individuate the referents of natural kind terms by their deep structure. That is, they should explore the possibility of giving up MPA. Secondly, they should try to find clearer cases than Twin Earth to support their thesis. Concerning the second option, it will clearly not do merely to change water for oil or cats for dogs. Some deeper rethinking of the approach is called for. I will pursue each strategy in turn, starting with the second.

There is a widespread assumption that externalism works only in depth. I think this is misguided. The most convincing counter-examples to internalism in fact involve surface-level differences between natural kinds. Take the following scenario.

7 That our cats and Mars’s robot cats share their properties above the molecular level does not mean that DNA is not part of what produces the superficial features of the former, as cells and organs do not last long without it. DNA stands in exactly the same relation to picobots as H2O stands to XYZ.

8 For instance, Schwartz (1979) asserts that Putnam is committed to the view that “at best the features associated with [a natural kind term] serve as good indicators of the presence of a real essence or nature that determines what it is to be a member of that natural kind” because he holds that natural kind terms are indexical. A noticeable exception is Putnam himself. In his 1992 paper he seems to consider the possibility of a more “superficial” externalism. It is unclear what he settles for, however. See Ayer’s reply in the same volume for a perspicuous characterization of his ambivalence.

9 David Hunter (2003) also makes the point that scenarios based on environments that differ in superficial features can constitute evidence for externalism. However, he fails to explain why this is so or provide a convincing scenario.
John takes part in a submarine expedition. Shortly after having reached the bottom of the ocean, he notices an octopus-like animal in the cameras. It appears to have ten tentacles and to be heavily carapaced. The species is baptized “philopus” in his honor (he is a philosopher). Some time after the dive, John learns that another John (John₂) who took part in another expedition in a different region claims to have discovered and baptized the philopus species before him. John₂ describes the specimen he observed exactly as John remembers the one he saw, and he wants to call the members of its kind “decapi.” Somewhat later, however, closer examination of the original specimens reveals that the one John observed has a feature the other does not have: its tentacles can regenerate. This turns out to be a very important characteristic of philopi, because they often have to give up a limb in their search for food in their inhospitable environment. We would clearly not say that “philopus” and “decapus” name the same species, yet at the time of introducing these terms the superficial characteristics known of either species were identical. This suggests that superficial features which play an essential role in grounding natural kind terms need not be known at the time of dubbing. This thought experiment provides evidence of a new kind for externalism because our judgment (or John’s) is not based on considerations of deep structure—it might well be that the structural properties that produce the superficial features common to decapi and philopi are exactly the same, and this does not seem to matter.

Let me give another example of this sort. As on Putnam’s original scenario, suppose that scientists travel to Twin Earth. They find that the watery stuff there is composed of XYZ. But they also find something else of interest. The space-travel technology they used to reach Twin Earth is powered by a new kind of field. It turns out that when exposed to this field XYZ becomes a powerful aphrodisiac, while H₂O has no such powers. Divergent intuitions have been expressed concerning Putnam’s Twin Earth, but the present scenario leaves little room for doubt: XYZ is not water. Our scientists on Twin Earth would be glad to announce that they have found a powerful aphrodisiac that looks just like water. On their side, Twin Earth scientists would be amazed by the discovery: something like “WATER IS AN APHRODISIAC” would make the headlines there.¹⁰ It is so clear that XYZ

¹⁰ The internalist who remains skeptical should give more unexpected powers to XYZ, e.g. that to burn like wood at temperatures and pressures never attained before the visit of Earthlings.
is not water in this scenario that it may seem that something illicit is going on. But what could that be? We fixed the extension of “water” centuries before meeting our counterparts on Twin Earth. We used to have exactly the same narrow\(^\text{11}\) beliefs as them about the watery substances of our respective planets, yet we were committed from the beginning to XYZ not being water.

The preceding scenario is more convincing than Putnam’s original because it involves differences in surface properties. But Putnam’s account of grounding does not take into account such “hidden” surface properties. It thus fails to explain the marked difference between the original Twin Earth and the aphrodisiac case. It also fails to account for the philopous case. In brief, it fails to capture the fact that externalism works in breadth—with surface properties that are not known at the time of dubbing. In addition, the text and robot cat scenarios suggest that externalism may not work in depth after all. While we have new, stronger evidence for externalism, we need a new account of grounding that fits all the data. I cannot provide a detailed solution here, but a rough sketch should be good enough for the purpose of shedding light on the philosophical doctrine of externalism.

We can probably retain Putnam’s idea that natural kind terms are introduced by ostensive definitions, although this is obviously a rough approximation of what really happens. For simplicity, let us assume that the extensions of natural kind terms are determined by ostensive definitions of the form: “‘T’ applies to whatever has enough of the important properties of this.” It will soon become clear why I added the qualification “enough”, which points to some subtle weighting of a kind’s important properties. All the difficulty is in the explanation of what important properties are. As just noted, we need to work on two fronts: we need to give more room to superficial features and less room to deep structure.

The so-called theory theory of concepts should be a good start. It may seem inappropriate to appeal to a psychological theory, but where else should we look for an account of the mechanism through which we fix the extensions of our natural kind terms? Even though part of it is in the world, it is clear that the aspects of this mechanism which currently elude us are in the head.

---

\(^{11}\) I invoke narrow beliefs because I take it that we have clear intuitions about narrow content. But here we need only say that we were in the same brain states as our counterparts on Twin Earth.
Also, a metaphysical account of natural kinds is of no use if it does not explain how we associate such entities with words. I thus concentrate on psychologists’ valuable contribution to the present problem, ignoring for the moment some excellent work in metaphysics.

The central idea of the theory theory is that the features of a natural kind which are essential are normally those which stand in explanatory relations to other essential features. Take the example of birds suggested by Murphy (2000). People think of birds as “feathered, two-legged creatures with wings, which fly, lay eggs in nests, and live in trees.” These are features which explain each other in folk ornithology: the power to fly explains the property of living in trees in conjunction with the need to stay out of the reach of predators; nesting is explained by the fact that eggs must not fall and that babies must have somewhere to stay before they are ready to fly; and so on. Keil (1992), Murphy and Medin (1999), Murphy (2000), and Kalish (2002) provide strong evidence to the effect that we categorize natural kinds on the basis of such theory-like clusters of mutually explanatory features. The point is not that laypersons have good or even important theories about natural kinds, but that these are thought to be good and influence their judgments of kind membership. These findings provide us with a good start, but we need be more explicit about the relation between the properties of a kind which are known at the time of dubbing and those which are not. I suggest the following definition of “important” features:

A feature of this object x is important if and only if
- it is an interesting, known feature of x,
- it explains an important feature of x,
- it is explained by an important feature of x.

Note that this definition is recursive. The first disjunct defines a base set of important features. The notion of interest that is relevant here is not technical. To illustrate, the specific shades of color found in the fur of cats are not interesting, whereas a philopus’ ten tentacles constitute an interesting feature. I specify that an instance of a kind needs only have enough of the important properties of the original specimen(s) to preempt the objection that many if not all the properties known a priori of most if not all natural kinds are not essential to them. (Remember that my concern here

---

12 I count as features both normal and dispositional properties (including causal powers).
is with the *grounding* of natural kind terms; it is clear that one need not know anything about a kind to refer to it if one defers to experts.) The two other clauses serve to expand the set of important properties associated with a kind beyond what is known of it at the time of dubbing. The expansion can proceed both in breadth and in depth, as both structural and superficial aspects of an object or substance can enter into explanatory relations with the rest of its features. Let me show how this is illustrated by the scenarios discussed above.

The philopus scenario is a case of expansion in breadth. I was careful to suggest that philopi’s regenerative capabilities are essential to them (in the practical sense of “essential”). This is why they matter for kind membership—not because these capabilities come with characteristic structural properties. Compare with another case. Not everybody had the chance to see a cat swim (or drown?), so let us suppose that the only difference between Earth and Twin Earth is that on Twin Earth cat-like animals explode when completely immersed in H\(_2\)O (there is H\(_2\)O on Twin Earth). This turns out to be due to small organic bombs they have in their tails. All Twin Earth cat-like animals have a gene which is entirely independent of the rest of their genome—which they share with our cats—and controls the creation of these bombs. It seems clear that in these circumstances the special feature of Twin Earth’s cat-like animals would not prevent us from calling them “cats.” At best we would say that there are two kinds of cats: the explosive and non-explosive ones. This contrasts with the philopus case, in which we want to say that philopi are not decapi, and vice-versa. The relevant difference is that there is no obvious explanation relating their bombs to cats’ way of life on Twin Earth, whereas there is an explanation that makes the special feature of philopi seem important.

The original Twin Earth thought experiment is also explained by the present account. Intuitively, the molecular composition of water on Earth explains its main superficial properties. It is thus important in the sense relevant to kind membership. Important properties are not automatically essential, so it may seem that my account has the unacceptable consequence that water is not necessarily H\(_2\)O. Not so. We must take into account the fact that we weight the important features of a kind differently depending on whether they are structural or superficial. That is, some features are more important than others. Individuals that greatly differ in their superficial features can still be members of the same kind, while individuals differing widely in
deep structure are usually not members of the same species. (See Gelman and Wellman [1999] for a good psychological study on this question.) In this sense, we have a bias for deep structure.

The proposed account makes the same prediction as Putnam’s concerning water in the actual world, but what explains the appearance of water in the text scenario is not that it has the deep structure E-A-U; it is the design and linguistic conventions of the creators. Consequently, its internal structure is unimportant and inessential, something MPA-based accounts cannot explain. The same holds of robot cats. DNA is relevant to the explanation of their most visible features in the actual world. It is thus important (and probably essential). But it would not be if the world turned out as in my robot cat scenario, because the existence of picobots would render it explanatorily irrelevant. All that would matter is that something or other regulates cat cells. It is clear that Putnam’s account fails on the text and robot cats scenarios because these bring out a crucial difference between causation (or production) and explanation, which is that the latter is more sensitive to background knowledge and interests.

The aphrodisiac case is a mixed one. The property of not being aphrodisiac stands in an indirect explanatory relation to water’s other interesting properties. It is not explained by or explaining water’s superficial characteristics, but it is explained by the same structural properties that explain its appearance. As a result, not being aphrodisiac is an important property of water. The aphrodisiac case is easier to judge than the original Twin Earth scenario because this indirect link gives us two reasons not to call XYZ “water”: it does not have the right deep structure, and it does not have the right surface properties.

My aim in this paper was to take what remains of the debate between internalists and externalists beyond a long-standing class of intuitions. The first thought experiments I considered illustrate this clash. My response is two-fold: scenarios that involve variations in surface characteristics provide stronger evidence than was previously available for externalism at the level of grounding, and an externalist account of grounding can accommodate the scenarios advanced by internalists. I call the view I sketched here “surface externalism”.

Surface Externalism
References