Pre-print of paper published in (2012) *European Journal of Philosophy*, 20(3), pp.389-406.

A Multiple Realization Thesis for Natural Kinds

Kevin Lynch

Abstract: Two important thought-experiments are associated with the work of Hilary Putnam, one designed to establish multiple realizability for mental kinds, the other designed to establish essentialism for natural kinds. Comparing the thought-experiments with each other reveals that the scenarios in both are structurally analogous to each other, though his intuitions in both are greatly at variance, intuitions that have been simultaneously well received. The intuition in the former implies a thesis that prioritizes pre-scientific over scientific indicators for identifying mental kinds in certain circumstances, while his intuition in the latter implies a converse thesis, prioritizing scientific over pre-scientific indicators for identifying natural kinds in analogous circumstances. In this paper I question whether we can consistently endorse both of these intuitions. A consideration is presented to attempt to justify the common intuition found in the multiple realization thought-experiment. Then it is argued that this same consideration has application in the structurally analogous Twin-Earth thought-experiment. This recommends a kind of multiple realization thesis for natural kinds, in opposition to a scientific essentialist approach. The various respects in which mental kinds like pain and natural kinds like water are similar to each other, such that similar philosophical treatments are warranted for both, are enumerated.

1. Introduction

The multiple realization (henceforth 'MR') thesis, often associated with Hilary Putnam, is the thesis that 'mental/psychological kinds', as with pain for instance, may be 'realized' by different kinds of physical system. The thought goes that octopi, extra-terrestrials or androids—who might very well be capable of feeling pain—could have their pain-states 'realized' by an underlying physical system radically different from ours in its material, structure and mechanics. A standard thought-experiment used to support this idea, derivable from Putnam's remarks (1967: 44), would go something like this. Imagine scientists discovered a certain physiological substrate to be necessary for pain experience in humans and the higher animals. Then we come across Martians (or octopi, or androids) who seem to feel pain like us (presumably because they display pain-behaviour like us), but who

at the same time have a radically different physiology. The intuition goes that we should take them to feel pain, despite their physiological variations. The case would only show that pain does not have a universally homogeneous substrate.

My interest here is with what this intuition (call it the MR intuition) implies about when there is a clash between the demands of different criteria for mental kinds (note that by 'criteria' I only mean properties which we take to *indicate* that something is an X, for example a creature in pain. I don't want to imply that these properties obtaining would *constitute the fact* that it is an X). In the thought-experiment there is competition between what I will call the 'pre-scientific', and also certain 'scientific' criteria for pain. These terms can be defined as follows. Firstly, with pain, there are the standard properties that we have always taken to indicate that a creature is in pain prior to any scientific investigations into pain's 'underlying nature', including characteristic behaviours occurring in response to injury. Call these the 'pre-scientific' pain criteria. And then there are the properties that, after scientific investigation, we found to underpin these more conspicuous reactions. These properties constitute the underlying physical basis or *substrate* of these pain experiences. We can call this substrate a 'scientific criterion' for pain: 'scientific' in that it was discovered by scientific investigation, and a 'criterion' insofar as we may use its presence or absence to tell if a creature is in pain. Our MR intuition then suggests a certain view concerning which criteria take priority in cases where there's a clash between the recommendations of each. When we judge that we obviously shouldn't deny that the Martians feel pain, what is suggested is that if we came across creatures that convincingly display the usual pre-scientific pain indicators, but not the scientific indicators as laid out in our best theory of pain-physiology, then the theory should capitulate to the pre-scientific judgments, not the other way round. Call this the Priority-of-the-Pre-Scientific-Criteria or PP-SC thesis.

In fact, Putnam would not endorse such a thesis without qualification, since in other papers he describes cases where the usual pre-scientific indicators of a psychological kind are absent, but where the substrate previously found to be associated with them is present, where this is supposedly evidence that the psychological kind is instantiated (see 1957 & 1963/1975). However, his MR arguments show the opposite pattern. He affirms that octopi 'certainly feel pain' (1967: 44), and so if their physiology were not of the same kind as ours,[1] this would only show that pain is multiply realizable. But how would he, or anyone else, have judged that octopi feel pain? Presumably on the basis of their capacity to display the pre-scientific criteria. He certainly wouldn't have judged it on the basis of physiological considerations, since we are to imagine octopi physiology to be different from ours, so in taking them to feel pain on that basis we would have to assume that this kind of substrate can realize pain, which would assume the truth of MR from the outset (the thought-experiment was supposed to demonstrate MR, not presuppose it).

It's likely that Putnam wasn't being inconsistent here, but that he thought the pre-scientific indicators have priority in some circumstances, but not in others. For example, it may be reasonable to take scientific indicators as trumping pre-scientific indicators with respect to certain individuals or groups, as with if Jones

were a criminal suspect taking a lie-detector test. This is because it is reasonable to assume that Jones is no lone exception to the rest of his species when it comes to the human physiological basis of lying, and plausible reasons can be given for not trusting his actions and words in these circumstances (he has a motive to deceive). However, such considerations can't be appealed to in the typical MR thought-experiment, where we are asked to imagine, say, that a *whole species* acts like we do when we are in pain, but doesn't have our physiology. A PP-SC thesis should be limited to such cases.

Whether Putnam held a PP-SC thesis or not, I will argue that we *should* endorse such a thesis (appropriately qualified to handle cases like that of Jones), and for a particular reason. This reason at least partly underlies and validates our MR intuitions in typical MR thought-experiments. My ultimate objective, however, is to show that the same consideration that validates the MR intuition conflicts with Putnam's intuition in his famous Twin-Earth thought-experiment about non-mental natural kinds, an intuition which forms the basis for scientific essentialism about such kinds. Thus it will be argued that those of us who share and endorse the MR intuition, and who are thus inclined towards MR, or the metaphysical possibility of MR, for mental kinds, should on pain of inconsistency also endorse an analogous thesis for non-mental natural kinds (I'll henceforth drop mention of the 'non-mental' qualifier).

One useful way of presenting the considerations which I believe to validate the MR intuition is by looking at what exactly goes wrong with taking the contrary position that would give priority to the scientific indicators for identifying instantiations of mental kinds. For this purpose I will discuss the views of the zoologist James D. Rose in section 2, who I believe exemplifies this strategy, also working with the case of pain. Then in section 3, I will explain why this approach goes wrong, such that a PP-SC thesis should be endorsed for pain. In section 4, I discuss Putnam's well-known thought-experiment about the natural kind, water. Though this, as I show, is structurally analogous to the MR thought-experiment, he displays markedly different intuitions, prioritizing the *scientific* criteria in this case. The intuition suggests a Priority-of-the-Scientific-Criteria (PSC) thesis for natural kinds, and the consequence of the thesis is that water can't have multiple realizations. Section 5 will concede that Putnam does allow for the possibility of multiple substrates for natural kinds in restricted circumstances. However, in section 6, the same argument that validates a PP-SC thesis for mental kinds will be mobilized for the case of natural kinds. In the final section, I enumerate the crucial similarities between mental and natural kinds which warrant similar philosophical treatments, and summarize the arguments.

2. Prioritizing Scientific Criteria

In recent writings, Rose (2002a; 2002b; 2007) claimed that it is impossible for fish to feel pain because they do not have the required physiological apparatus. In the case of humans and other primates, the theory is that on injury, nociceptive receptors

send signals for processing to the neocortex, terminating in the experience of pain. Although nociceptive receptors of some form may be present in fish (excluding sharks), fish do not have a neocortex, and for this reason they cannot, in Rose's opinion, experience pain (or have any conscious experience for that matter).

Rose's assertions raise issues on the relationship between pre-scientific and physiological criteria for mental kinds, and make us wonder under what circumstances we should take one to trump the other. He takes a view on what should guide our decision-making for ascribing pain and consciousness in the case of fish. He implies that once the physiological substrate of pain is identified in higher animals, its presence or absence in these lower life-forms may serve as a criterion for telling if they can experience pain or conscious states.

According to Rose, the fundamental error that we make in thinking fish could feel pain is conceptual and methodological (2007: 139). It arises from our 'anthropocentrism': 'in observing the actions of other organisms where the actions appear to resemble our own, often it is assumed that these non-human organisms have intentions and experiences similar to ours' (Rose 2002b: 3, also see Harrison 1991: 35–36). This assumption, he maintains, has led to error by those who have observed pain-behaviour in fish similar to human pain-behaviour. For example, it has been observed that the avoidance behaviour of fish to noxious stimuli involves more than mere reflex responses. Classical conditioning experiments have shown that fish steer clear of harmless stimuli that they learn to be associated with an aversive stimulus (Topál & Csányi 1999; Yue, Duncan and Moccia 2008). Also, Lynne Sneddon (Sneddon 2003; Sneddon, Braithwaite and Gentle 2003) and her colleagues have found that trout injected with acid or venom into their lips will sometimes rub their lips on the fish-tank wall or ground, will temporarily neglect normal activities such as eating, and will engage in anomalous behaviour like rocking, all of which they think resembles, to some degree at least, how certain primates respond to noxious stimuli (think of rubbing the area of a nettle sting to sooth the pain). Opiates were found to reduce these behaviours. She takes these findings as supporting the claim that fish feel pain. In Rose's view, however, we would be mistaken in concluding on such grounds that fish actually feel pain (and that they 'fear' a pain-causing object when they avoid it), or are conscious, on the basis of these responses, since such complex behaviours can occur in humans in the absence of consciousness (Rose 2007: 145–146).[2] What is decisive, in Rose's view, is whether fish have the apparatus discovered to cause pain in humans and the other higher animals. Once science has found this property, its presence or absence may overrule the presence of absence of behavioural indicators. Rose would find some ardent allies among philosophers. John Searle would endorse the privileged role of this 'scientific criterion':

> Regardless of how their behaviour differs from human behaviour, oysters might still be conscious if they had the right sort of neurobiological processes in their nervous systems. Suppose we had a perfect science of the brain and we knew exactly how consciousness is produced in humans and the higher animals. If we then found that the consciousness-producing

> mechanism was present in oysters but not present in snails we would have very good reason, indeed, overwhelming reason, for supposing that oysters are conscious and snails are probably not conscious. (Searle 2007: 104. Also see Searle 1992: 74–75 & Searle 1994: 206–219)

In Searle's opinion, if we discover some neural feature *N* which is necessary and sufficient for consciousness in humans and the higher animals, then the presence or absence of *N* in a species whose behaviour and morphology is too ambiguous to support a confident decision will help us decide if it has consciousness. Not only that, Searle's statements imply that the presence or absence of *N* in something could *overrule* the ascriptions we would tend to make on the basis of behaviour/morphology. This seems safe to say since Searle thinks that its presence in *oysters* would indicate that they are conscious, and we would surely be inclined to deny on the basis of behaviour/morphology that oysters are conscious (since they are amorphous blobs with no sense-organs, and are capable of no behaviour at all).

The challenge for someone taking Rose's position would be to explain why we should not take the physiological facts to mean only that pain in fish has a different physical realization than in primates. Putnam, I gather, would think that Rose is getting things backwards. Using the examples of reptiles, octopi and aliens (1967: 44–45), he seemed to imply that if our pre-scientific judgments that they can feel pain were in opposition to our best physiological pain-theory stating that all creatures capable of pain must have neural apparatus N, then these judgments would overrule the theory. Though he doesn't talk about fish, one might assume that there is no great leap from octopi to fish. The different physiologies of these creatures would show that pain has no universally homogeneous substrate.

This is the MR thesis, and it was left by Putnam on an intuitive level. I now wish to show how taking the contrary position would actually be illogical, by showing what is illogical about Rose's position.

3. Discovering the Substrate of Pain

Against the view that fish feel pain, Rose does adduce behavioural evidence showing remarkable resilience to injury in fish, though one could argue this only suggests that their sensitivity for pain is not as robust as with humans. But physiology is the decisive factor for Rose. His position is that fish can't feel pain on the grounds that they lack a neocortex, and so he must be assuming that possession of a neocortex is a necessary condition for pain experience. It is not clear whether he would maintain this if we met human-like aliens without a neocortex, but it's hard to see what *principled* reasons he would have for relying on 'anthropomorphic' criteria with the aliens, though not for the fish. So let's assume that he does think that the neocortex is essential.

The question is: how would this have been discovered in the first place? It seems that in order to find out what physiological equipment is required for the

experience of pain, a *representative sample* of subjects would first have to be selected (human, or monkey perhaps) to find out what goes on inside them when they are in pain. How else could such an investigation get started? But in selecting this sample, we would have to already know that they can experience pain. How could we know that? Don't say we could only know that by seeing what internal condition their body is in, because that is exactly what has yet to be established: what internal condition they must be in to experience pain. The selection criteria for subjects for the tests must have nothing to do with whether they have a neocortex, or noiceceptors, or c-fibres, or even a brain, because the scientists as of yet don't have a well-tested theory of pain physiology. The selection criteria must be based on whether they can display the pre-scientific criteria which the scientists rely on in their capacity as ordinary members of our linguistic community. In principle there could not be a verified physiological theory of pain, if scientists did not originally accept the pre-scientific criteria as indicating that certain creatures can experience pain.

Though it is not necessary for our purposes to specify exactly what make up these criteria, one question worth addressing is whether *self-reporting* behaviour, generally the reserve of humans, is the primary, most reliable, or most direct indicator for pain, compared to the other more primitive pain-behaviours we share with other animals. For if self-reports are primary, then doubts might be raised about whether animals would be appropriate candidates for our sample. However, there are some strong reasons for thinking that self-report is in an important sense a derivative criterion, reasons that may be gathered from some relevant remarks by Ludwig Wittgenstein. One of these arguments starts by asking us to consider how the word 'pain', which we use in these self-reports, could have got its meaning. Now if it were the case that 'pain' got its meaning by my associating it directly with the subjective experience of pain, then my making the self-ascriptions 'I'm in pain' would indeed be a particularly direct form of evidence that I really am in pain. It would also be a form of evidence for pain that's completely independent of the more primitive behavioural indicators.

However, 'pain' can't have got its meaning in this way (Wittgenstein 1953: §293). For if this is the way that 'pain', as I use it for self-ascribing, got its meaning, then what is to ensure that what *I* have associated 'pain' with is the same as what *others* associate 'pain' with? What is to ensure that 'pain' refers, for all members of the linguistic community, to the same thing? On the view that 'pain' gets its meaning by being associated by each of us, through a private act of ostensive definition, directly with a subjective experience, there would not yet exist a public concept of pain, only a bunch of private ones (though Wittgenstein has something to say against the possibility of even these private concepts). So for 'pain' to have a *common meaning* that can be of service in communication, 'pain' must have got its meaning by being associated with something publically observable, something we can all together associate the word with, the likely candidate being these more primitive pain-behaviours, and the circumstances in which they occur. That is to say, we stipulate, and then learn and teach the proper use of the word 'pain' with reference to these behavioural displays. This would

imply that self-reports are not a more fundamental form of evidence for pain than these more primitive pain-behaviours. So if 'pain' got its meaning by being associated with certain publically observable criteria (behavioural etc.), then self-ascriptions employing that word can't (in general) be a more direct or certain form of evidence for pain than these more primitive criteria, nor can they be a source of evidence fundamentally independent from these criteria.

Turning again to our scientists' original endeavours to discover the substrate of pain, imagine now that they discover the neocortex plays an essential role in causing pain for the sample of subjects they have studied. Since this sample will only ever be a small subset of the set of possible samples (they can't test every relevant person/creature), in constructing a general theory of pain they must *generalize*, saying that all other subsets that would have served just as well as representative samples would have had this internal apparatus too. Now say that afterwards a creature is found that's capable of the responses which constituted the selection criteria for subjects used in the original tests, but that it has no neocortex. If we maintain that this creature cannot possibly feel pain we are committing an error, an inconsistency. We may realize that *had it been included in that original sample* used in the original investigations, then we would never have ended up making the generalization that the neocortex is necessary for pain experience. In that case, it looks arbitrary when we claim that it could not experience pain on the basis of the scientifically irrelevant fact that it happened not to have been among those subjects who were randomly included in the sample when the scientists began investigating the physiology of pain, though it would have qualified for acceptance since it can display the sample selection criteria.[3]

A similar error is implicit in Rose's work. The selection criteria for subjects for the physiological studies can only be based on their capacity for responses stereotypically associated with cases of people being in pain (which may include anything from involuntary responses like grimacing, to rational/intentional responses like avoiding the source of the injury or self-ascribing pain). There would have been no reason to deny any individual admittance into the sample of test subjects that could display these properties. So any kind of normal human being would have been admissible into this sample. Other primates would also have been admissible (as Rose notes, they have been routinely used in investigations into the neural basis of 'primary consciousness' [2002b: 14–15]). And might we claim that *fish* could have been admissible too? I don't want to argue that fish would have been admissible; all I want to claim is that the grounds on which we would or would not have admitted fish could only have been based on their visible reactions and responses, not physiological considerations, because excluding them on physiological grounds would presuppose the correct physiological theory yet to be established. So the key question to ask is: does this thing satisfy the selection conditions that would have admitted it—had it been in our vicinity and available back then—into our subject sample when the scientists were collecting samples to investigate the physiology of pain?

Even Rose's claim that the neocortex is necessary for pain *in humans* seems very dubious. He says that '[o]ur fundamental behavioural reactions to noxious

stimuli, including vocalization, facial grimacing, and withdrawal, are mediated by subcortical brain and spinal systems. Activation of these responses by noxious stimuli can occur without consciousness in people with extensive cortical damage and in humans born without cerebral hemispheres' (Rose 2002b: 17). So people born without a neocortex can still display what Rose seems to imply are the usual primitive pain-responses.[4] Yet it may very well have been the case that these exact responses, vocalization, grimacing, withdrawal etc., often made up the selection criteria for the samples during the original scientific investigations, which is presumably true if monkeys and even more primitive animals like rats have been sometimes used in such studies, as is the case (for a review see Rainville 2002). If those people born without a neocortex had been around when the scientists were looking for test-subjects, couldn't they have been included in the test-sample? Certainly they couldn't have been excluded on the grounds of them having no neocortex, because the scientists didn't have a verified physiological theory then detailing the neocortex's role.

The following principle is suggested by the preceding remarks:

> *It would be unwarranted to exclude anything as a possible counter-instance to a generalization if it would have counted against that generalization being made, had we known about it before we made the generalization.*

This I put out in the open as the basic premise of this paper, which I will call the Consistency Principle. I suggest that it at least partly explains and validates our intuitions in the MR thought-experiment (and later I will argue that it legitimates similar moves for natural kinds), and I admit that if the reader doesn't find this principle as intuitively compelling as I do, I probably won't be able to convince him/her of this paper's thesis.

4. Putnam's Different Approach to Natural Kinds

Putnam is also a well-known founder of the theory of scientific essentialism (SE) about natural kinds. According to SE, a natural kind (such as a kind of animal, mineral, or liquid) can't have multiple realizations/substrates because it's individuated with respect to a specific kind of physical substrate, a substrate that is the 'essence' of that kind. This, then, would mark a fundamental difference between natural and mental kinds. His argument for SE was based on a thought-experiment, and an intuition (1975; 1994).

The thought-experiment can be abridged as follows. Imagine that we find, as we have in Putnam's view, that water on Earth is H_2O at the molecular level.[5] Then imagine that we later find on another planet (called 'Twin-Earth') a substance which has the exact same, or near enough the same macroscopic properties as water on Earth (properties which either are or at least include the properties used pre-scientifically to identify a substance as water), but which does not share the microphysical structure of Earth-water (its chemical formula is something other than H_2O, represented by 'XYZ'). Putnam's intuition is that this

substance would not be water. Water is simply identical with the molecular compound H_2O.

Though it may indeed be chemically impossible for this Twin-Earth scenario to arise, the assumption here is that it is conceivable (conceivability does not seem constrained by physical impossibility). Or at least Putnam seems to be assuming so when he asks us to imagine such a world in his thought-experiment. Putnam thus follows the venerable philosophical tradition of inventing conceivable though perhaps physically impossible scenarios to elicit our intuitions about what we should say in that situation, which is supposed to tell us something about the concepts involved. Note that Putnam offers no *further* argument explaining *why* we should agree that this liquid would not be water. It is supposed to just strike us that it obviously would not. However, intuitions have actually been divided over this, much more so than over the MR intuition it would seem. Some declare his proposal as 'natural and obviously correct' (Burge 1996: 144). Others have found it less compelling, and argue that it's at odds with historical linguistic practice (e.g. Zemach 1976; Cassam 1986; Smith 2005). Smith reports that in his over 20 years of teaching the Twin-Earth thought-experiment to students, 'the overwhelming majority of them regard it not only as true, but as obviously true, that twin-water is water' (2005: 81). On the other hand, a recent formal study probing the intuitions of two groups of students found a majority to have essentialist intuitions in a Twin-Earth style case (Jylkkä, Railo and Haukioja 2009). On the whole, ordinary intuitions seem somewhat unstable on this issue, and it's unclear, at present, what to make of all these divergent claims.

Be that as it may, Putnam's intuition seems to imply a Priority-of-the-Scientific-Criteria (PSC) Thesis for natural kinds, at least for circumstances related to that of the thought-experiment. In this case, there was a dissociation between pre-scientific and scientific indicators. The 'manifest properties' of the XYZ liquid constituted or included the standard 'pre-scientific criteria' for identifying water, while the 'scientific criterion' for water—H_2O—was absent. Given that we are to conclude that this liquid is not water, this suggests that the scientific indicators have priority over the pre-scientific.

What I want to highlight is how different this SE intuition is from the MR intuition, a point that has been noted by Bealer (1994: 199). This is surprising considering the close structural affinities between the MR and SE thought-experiments, and also considering how Putnam probably thinks of pain as a 'natural kind' in some sense also. Let me lay out their structure side-by-side (see Box 1) in order to make the affinities, and also the differences between intuitions in these thought-experiments, appear more vivid (note that these thought-experiments may not have been formulated exactly in this manner by Putnam, though these formulations can easily be derived from his remarks).

The fact that the received intuitions here differ so starkly, despite the structural similarities between the thought-experiments, has stimulated surprisingly little discussion. This is not to say that philosophers haven't made comparisons between propositions like 'water is H_2O' and, say, 'pain is c-fibers firing'. Far from it, many philosophers have indeed *emphasized* the analogousness of the two

Box 1. The thought-experiments

	MR thought-experiment
Scenario:	1) Scientists discover that pain-responses in pain-experiencing life-forms on Earth are underpinned by a single type of physiological substrate. 2) A life-form is encountered not from Earth which displays the same responses, but with a different substrate.
Intuition:	This off-world entity *would* be an instance of the original kind: pain-experiencing life-form.
Conclusion:	Pain is multiply realizable.
	SE thought-experiment
Scenario:	1) Scientists discover that the manifest properties of water on Earth are underpinned by a single type of molecular substrate. 2) A substance is encountered not from Earth which displays the same manifest properties as Earth water, but with a different substrate.
Intuition:	This off-world substance *would not* be an instance of that original kind: water.
Conclusion:	Water is not multiply realizable.

cases, in order, for instance, to allay qualms about the intelligibility of identifying pain with a brain state (it is said to be *no less* intelligible than identifying water with H_2O, see, for example, Putnam 1967: 39). They have, indeed, tried to model the latter kind of fact on the former. But this makes it all the more surprising that the contrariety of the above intuitions have attracted such little notice or puzzlement, as if there were no need to explain why they should be taken as simultaneously valid. However, in light of the mentioned structural similarities, some such explanation would surely be needed. Presumably one would have to show that pain is disanalogous to water in some crucial respect that explains and justifies our markedly different intuitions in the two cases.

It is worth mentioning here Kripke's (1980) comparison of propositions like 'pain is c-fibers firing' and 'water is H_2O', since he claimed to have found profound disanalogies between them. However, as I would like to emphasize, Kripke comes at this issue from a very different angle from what we have been looking at. Kripke compares the way we 'pick out' pain with the way we would 'pick out' a natural kind (he uses the example of heat, but let's stick with water here). However, in the former case he is thinking of picking out pain *in ourselves*. He then argues that while in the latter case, we ordinarily pick the thing out through acquaintance with what he takes to be its 'accidental properties' (the macroscopic characteristics of water), in the former we don't, since we pick it out by feeling the pain itself (and to feel pain is to be in pain).

However, in the MR thought-experiment above, we are imagining trying to 'pick out' pain *in a being other than ourselves*, and we are comparing this with a situation of trying to identify water. Because we are trying to identify pain in another, we must rely on pain-indicators or criteria (verbal or non-verbal behaviour). This, however, we *don't* need to do when identifying pain in ourselves. We don't need to observe our own pain behaviour, or observe ourselves self-ascribing pain, to know that we

are in pain. We can 'pick out' the pain 'immediately' by feeling it, without relying on any criteria at all. So the comparison above is coming at things very differently from Kripke's. It is specifically concerned with the issue of identifying a being *other than ourselves* as one that can or cannot feel pain, where all we have to rely on is pre-scientific criteria (verbal or non-verbal) or scientific criteria, and it disassociates these criteria to see how our intuitions go, where this is then compared with a case of trying to identify a substance as being or not being water, where all we have to go on are the pre-scientific criteria (the observable behaviour/properties of the substance), or its scientific criteria. It is only when we think of picking out pain in the third-person case that the analogies between pain and water emerge. I hope I am right in thinking that the fact that the analogy doesn't hold when we think of identifying pain in our own case does not undermine the fact that the analogy between the cases does hold when we consider identifying pain in others, and compare this with trying to identify a natural kind.

The similarity between the two cases puts the onus on the essentialist to explain why we should go one way with natural kinds and the other with mental kinds. I, however, will not be trying to defend the essentialist intuition, but will try to make a case for giving a consistent treatment to both.[6] It will be argued that the similarities between mental and natural kinds are such that the same considerations that justify a PP-SC thesis for mental kinds are applicable to natural kinds also. But first I wish to concede that Putnam does admit the possibility of multiple realizations/substrates for natural kinds within limits.

5. Putnam's Limited Admission of Multiple Realizability for Natural Kinds

Putnam and Kripke hold that the reference of a natural kind term is determined by the underlying nature of a *particular sample-set* of the kind which the users of the language are acquainted with or in proximity to. We may call this the *determining set*. How they conceive of this determining set as being circumscribed, however, is not fully clear, and seems to vary. Putnam vaguely identifies it with the 'local' samples 'around here' or in 'our environment'. At other times it's suggested that the determining set is the samples referred to in the initial act of ostensive definition (if there was one) when the term was introduced. Or sometimes it's suggested that it's the entire collection of samples on Planet Earth, or the samples originally gathered by the scientists for analysis. What all these positions would have in common, however, is the idea that the reference gets 'fixed' onto the underlying nature of a particular set of examples. Whatever *their* underlying nature is, the term's meaning is fixed 'rigidly' to that. Putnam calls this the 'indexicality' thesis.

However, in relation to this determining set, there is *no a priori guarantee* that the things in it will have a common nature:

> It could have turned out that the bits of liquid we call 'water' had *no* important common physical characteristics *except* the superficial ones. In that case the necessary and sufficient conditions of being 'water' would

> have been possession of sufficiently many of the superficial characteristics . . . If H_2O and XYZ had both been plentiful on Earth, then we would have had a case similar to the jadeite/nephrite case: it would have been correct to say that there were *two kinds of 'water'*. And instead of saying that 'the stuff on Twin Earth turned out not to be really water', we would have to say 'it turned out to be the XYZ *kind of water* . . . To sum up: if there is a hidden structure, then generally it determines what it is to be a member of the natural kind . . . But the local water, or whatever, may have two or more hidden structures—or so many that 'hidden structure' becomes irrelevant, and superficial characteristics become the decisive ones. (Putnam 1975: 241)

Here Putnam implicitly takes the samples on Planet Earth, samples of 'local water', as the determining set. Kripke admits something similar, taking what he calls the 'original samples' to be the determining set. By 'original samples' I think he means the samples used in the first scientific analyses to determine their underlying nature, as opposed to the samples referred to in an original ostensive definition of the kind term (the difference is not philosophically significant I believe).

> If the original samples [of gold singled out for investigation into their structure] have a small number of deviant items, they will be rejected as not really gold. If, on the other hand, the supposition that there is one uniform substance or kind in the initial sample proves radically in error, reactions can vary: sometimes we may declare that there are two kinds of gold, sometimes we may drop the term 'gold'. (These possibilities are not supposed to be exhaustive.) (Kripke 1980: 136)

The view here seems to be that if in the *actual world* we discover the determining set of water samples to consist of H_2O, then the meaning of 'water' gets fixed rigidly to that, so that if we then were to come across a Twin-Earth like liquid on some other planet sometime in the future (which although perhaps physically impossible, is yet conceivable) then this substance would not be water. Thus the claim is not that water is necessarily H_2O in all conceivable possible worlds (since in some worlds, the watery stuff in our oceans etc., might not have been H_2O), but only that it is necessarily H_2O *given* that the actual world is one in which the determining set turned out to consist of H_2O (on this, see Chalmers 1996: 57–59).

We also see from this quotation how Putnam's theory allows for the possibility of multiple realizability for natural kinds in certain circumstances, namely, where the things in the determining set don't share a common substrate. According to the indexicality thesis, a natural kind term's meaning is fixed by the nature of a crucial determining set, a set of 'local' samples. *If* everything in this set is discovered to share a common substrate, then the term rigidly denotes that. But there is no guarantee that the samples in this determining set will have a common substrate. Putnam offers jade[7] as an example of a multiply realized natural kind,

because here the determining set (i.e., samples satisfying certain pre-scientific criteria on Earth) consisted of two different compounds. But if all the jade on Earth had been nephrite, then 'jade' would have rigidly designated nephrite, since all the samples in the determining set (all 'local samples') would have shared that one substrate. In such circumstances, were we to then find jadeite on another planet, this would *not* be jade.

6. Running the PP-SC Argument for Natural Kinds

So Putnam does give a restricted admission that MR is possible for natural kinds. However, scientific essentialists still give the scientific criteria a questionable primacy in circumstances where the things in the determining set happen to share a common substrate. To see this, let's switch to the example of nephrite and jadeite instead of H_2O and XYZ (since thereby our thought-experiment will have the virtue of not involving any elements incompatible with the actual laws of nature, as may be the case with Putnam's XYZ liquid). Consider the counterfactual world mentioned above where all the jade on Earth happens to be nephrite (where 'nephrite' represents a complex chemical formula). Here 'jade' would rigidly designate that compound. Now in this counterfactual world, when the samples of jade were first gathered to investigate its microphysical nature, they had to be gathered in accordance with the pre-scientific criteria.[8] No selections or exclusions could have been made on the basis of scientific criteria, since no well-tested scientific theory yet existed. The samples were analyzed and it was discovered that they all consisted of nephrite, and it was declared that jade is nephrite. But the scientists were not interested in *those* samples *as such*; they were interested in them as *representatives* of the kind. Accordingly, their declaration amounts to a *generalization* across all substances that would also have qualified as a representative sample, because they satisfy the sample selection criteria (one might claim that they don't *generalize*, but that they *stipulate* on the basis of those experiments, that all jade is nephrite, but the truths of mineralogy are hardly safe-guarded by fiat).

Afterwards a substance is found on another planet displaying (near enough) the properties that constituted the selection criteria used for gathering test-samples of jade on Earth, but with a different substrate (represented by our term 'jadeite'). But by Putnam's view, this substance would not be jade, because it doesn't have the substrate of Earth jade. However, had this substance been around on Earth when the scientists were collecting their representative samples, they would have selected it, and then they never would have made the generalization that all jade is nephrite in the first place. This follows from Putnam's admissions in the cited passage. So the essentialist wants to exclude it from the category 'jade' just because it happened to be 'over there' instead of 'around here' when the tests were being done. This is what makes the moves of the essentialist problematic, in a way exactly analogous to the moves made by Rose regarding pain. Rose excluded creatures with no neocortex from the category of 'creatures capable of experiencing

pain' even though they would have, in all likelihood, prevented the generalization 'all creatures capable of experiencing pain have a neocortex' from ever being made had they been picked when the scientists were collecting subjects for their experiments (and there's no reason why they shouldn't have been picked if they displayed the sample selection criteria). If our acknowledgement of this as problematic leads us to endorse a PP-SC thesis for mental kinds, why should it not do the same regarding other natural kinds?

Another way of looking at what is troublesome about having the question of whether something does or does not belong to the categories of 'jade' or 'creature that can feel pain' depend on the scientific criteria is that this means that the question may turn on what its position in space and time is, rather than on its intrinsic nature. In our counterfactual world, had jadeite been on Earth it would have been classified as a kind of jade. But because it happened to be on another planet, it doesn't get classified as jade. However, spatial location just does not seem like a pertinent consideration here. One would think that such things should depend just on the intrinsic nature of the substance.

7. The Similarities between Mental and Natural Kinds

Let me end by enumerating the important respects in which a mental kind like pain, and a natural kind like water or jade, are similar. Then I will summarize the argument for why we should endorse MR for both kinds. The crucial similarities are as follows:

(1) Both are phenomena that have underlying physical *substrates*, such that the scientific/pre-scientific criteria distinction, in the present sense, applies.

(2) When we seek to discover the hidden substrate of both, we first must conduct scientific studies that involve *selecting representative samples*.

(3) In both cases, those samples cannot be gathered on the basis of criteria advanced in a scientific theory, because what the underlying nature is which science seeks has at that stage not been established. They must be *gathered on the basis of the pre-scientific criteria*.

(4) In both cases, the samples may be subjected to analysis, and it may or may not be found that in the relevant respect, the members all have a common underlying nature N.

(5) If they do, then on the basis of these studies, we may feel confident enough to pronounce that 'the underlying nature of the mental/natural kind is N.' Because the sample was one of many possible legitimate samples (that is, it's a *representative* sample), this pronouncement *constitutes an implicit generalization* (i.e. it embodies the presumption that the many other possible legitimate samples they might have analyzed would have had N too).

Then, on an SE approach, the proposition that the underlying nature of the kinds is N is taken to embody a necessary truth, *such that N takes criterial priority over the original sample-selection criteria/pre-scientific criteria* (as Putnam explicitly does for water, and as Rose does in practice for pain). Given the mentioned similarities, this approach can be seen as problematic because it leads to the following reasoning:

(a) *Ex hypothesi*, substrate *N* is essential for *K* (where *K* is a mental or natural kind).
(b) An example *E* may then be encountered without underlying nature *N*, but which displays (to a near enough degree at least) the pre-scientific criteria that served as the sample selection criteria during the original investigations into *K*'s underlying nature.
(c) Because of (a), *E* is not regarded as a genuine example of *K*. The absence of *N* overrules the presence of the pre-scientific criteria.
(d) Yet had *E* been available during the original investigations into *K*'s substrate, and before *N* was discovered, it would have been accepted into the representative sample (having the properties that constitute the sample selection criteria). And then it would have prevented the generalization '*K* is *N*' from ever being made.

These steps show a problem because they breach the aforementioned Consistency Principle: if something would have prevented a generalization from being made had we known about it before making that generalization, then its discovery afterwards should be counted as undermining that generalization, no matter when, where, or in what quantities it is discovered. I am assuming that this principle itself is in need of no argument. It is just a plea for consistency in how we would treat E. Accepting this, we should reject the SE intuition and welcome the MR intuition for *any* kinds that are suitable for taking the place of the variable '*K*'. The discovery of an E should always lead us to conclude that *K* can have realizations other than *N*. Of course, someone defending Putnam's disparate treatments might describe respects in which mental and natural kinds are *dis*similar, such that one might be justified in being a substrate essentialist about natural kinds, and an anti-essentialist MR proponent about the mental. I'm sure they are dissimilar in many respects, but the point is that they are similar with respect to points (1) to (5), which should be sufficient for showing that taking the SE approach will lead to transgressing the Consistency Principle. And that should be enough to detract from the SE approach for those kinds.

None of this is to deny that discovered substrates can be very useful as identifying criteria. But this is mainly because what Putnam calls 'manifest properties' are not always so manifest. Dispositional properties, for instance, are plausibly among the pre-scientific criteria of natural kinds. But we can't witness these properties by a quick examination. If we went to a new planet and saw stuff that seemed identical to water, perhaps it's actually poisonous. Rather than drinking it and finding out the hard way, better test it. If it's H_2O we can be reasonably confident it's safe. If we have doubts about a purported piece of gold,

why wait around to see if it tarnishes after a month when we've lost the receipt, when we can appeal to a scientific criterion? The lie-detector test is another example.

My present aim has been to validate the MR intuition, an intuition which is at least partly based on a requirement of consistency regarding when something is to be taken as a counterexample to a generalization. I have tried to show that these same considerations may legitimately apply to natural kinds, and if so, the SE intuition should be rejected. If one was so inclined, one could take these thoughts in the direction of a global functionalism. On this view, our metaphysical view of water would be analogous to a functionalist metaphysical perspective on pain; water would be defined as that substrate (or those substrates) which has the function/role of causing a substance to have the higher-level properties or 'outputs' pre-scientifically associated with water. I, however, am not sure that MR entails functionalism, so that would need further argument. MR is just a possibility that arises once we accept the priority of pre-scientific over scientific indicators for kinds which have substrates.[9]

Kevin Lynch
Department of Philosophy
Warwick University
UK
kevinlynch405@eircom.net

NOTES

[1] The pain-physiology of octopi would have to be judged to be of a 'different kind' to that of humans with respect to some valid criteria for physiological kind individuation. There currently is debate over what the appropriate criteria should be, and some of the options would be more congenial to the possibility of MR than others. I have doubts about whether there is a single, privileged way of individuating physiological kinds, but for our purposes let's just assume that relative to some plausible criteria, octopus pain would come out as having a 'different physical realization' to our pain.

[2] Actually, this point is debatable. Rose refers to cases of complex behaviour performed during sleepwalking. But sleepwalking might be considered a semi-conscious state. Though not like full wakefulness, it is far removed from paradigms of unconsciousness like being comatose.

[3] Similar points are made in relation to using a physiological criterion as the primary indicator governing our ascriptions of consciousness in Chalmers 1998. Also see Morick 1971: 292–295.

[4] On the issue of people born without a neocortex and their capacities, see Merker 2007.

[5] We may ignore, for argument's sake, criticisms that this claim may be an over-simplification (e.g. Zemach 1976). Complications arise here due to the fact that normal samples of water generally contain other elements besides H_20 that contribute to its macroscopic properties, and also concerning the fact that isotopes of hydrogen can occur, such as deuterium, which is present in D_2O (commonly called 'heavy water'), which behaves similarly to H_20.

[6] Kim (1992) has also tried to draw parallels between the cases of pain and the natural kind, jade, saying that whether we decide to regard them as disjunctive kinds or as not kinds at all, they should be treated similarly.

[7] It was discovered in the nineteenth century that what people took to be the one kind of ornamental stone, jade, could be either of two different mineral compounds, nephrite or jadeite. These are macroscopically very similar (though not exactly so).

[8] This fact may have motivated Chalmers (1996) and Jackson (2000) to associate a 'primary intention' with natural kind terms in their two-dimensionalist analysis of natural kind term meaning. What I call the 'sample selection criteria' would be associated with the primary intention on their view. Chalmers and Jackson, however, accept Putnam's intuition in the Twin-Earth thought-experiment (which in turn motivates them to associate a 'secondary intention' with natural kind term meaning, constituted by the substrate of the determining set), the validity of which I dispute in this paper.

[9] Shorter versions of this paper were presented in 2009 at the Metaphysics of Mind conference, Edinburgh University, and at the British Postgraduate Philosophy Association conference, King's College, London, and earlier at a work-in-progress seminar at Warwick University, and I thank the audiences for their comments. Thanks are also due to Johannes Roessler and to an anonymous referee from this journal for very useful feedback on earlier drafts. I'm also very grateful for a generous scholarship from the Department of Philosophy, Warwick University, which supported this work.

REFERENCES

Bealer, G. (1994), 'Mental Properties', *Journal of Philosophy*, 91: 185–208.

Burge, T. (1996), 'Other Bodies', in A. Pessin and S. Goldberg (eds.) *The Twin Earth Chronicles*. Armonk, New York: M.E. Sharpe.

Cassam, Q. (1986), 'Science and Essence', *Philosophy*, 61: 95–107.

Chalmers, D. (1996), *The Conscious Mind*. New York; Oxford: Oxford University Press. —

— (1998), 'On the Search for the Neural Correlate of Consciousness', in S. Hameroff, A. Kaszniak and A. Scott (eds.) *Toward a Science of Consciousness II*. Cambridge, MA: MIT Press.

Harrison, P. (1991), 'Do Animals Feel Pain?', *Philosophy*, 66: 25–40.

Jackson, F. (2000), *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Jylkkä, J., Railo, H. and Haukioja, J. (2009), 'Psychological Essentialism and Semantic Externalism: Evidence for Externalism in Lay Speakers' Language Use', *Philosophical Psychology*, 22: 37–60.

Kim, J. (1992), 'Multiple Realization and the Metaphysics of Reduction', *Philosophy and Phenomenological Research*, 52: 1–26.

Kripke, S. A. (1980), *Naming and Necessity*. Oxford: Basil Blackwell.

Merker, B. (2007), 'Consciousness Without a Cerebral Cortex: A Challenge for Neuroscience and Medicine', *Behavioral and Brain Sciences*, 30: 63–134.

Morick, H. (1971), 'Is Ultimate Epistemic Authority a Distinguishing Characteristic of the Psychological?', *American Philosophical Quarterly*, 8: 292–295.

Putnam, H. (1957), 'Psychological Concepts, Explication, and Ordinary Language', *Journal of Philosophy*, 54: 94–100.

—— (1963/1975), 'Brains and Behavior', in *Mind, Language and Reality, vol.2*. Cambridge: Cambridge University Press.

—— (1967), 'Psychological Predicates', in W. H. Captain and D. D. Merrill (eds.) *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press.

—— (1975/1975), 'The Meaning of ''Meaning''', in *Mind, Language and Reality, vol.2*. Cambridge: Cambridge University Press.

—— (1994), 'Is Water Necessarily H_2O?', in J. Conant (ed.) *Realism with a Human Face*. Cambridge, MA; London: Harvard University Press.

Rainville, P. (2002), 'Brain Mechanisms of Pain Affect and Pain Modulation', *Current Opinion in Neurobiology*, 12: 195–204.

Rose, J. D. (2002a), 'Do Fish Feel Pain?', http://cotrout.org/do_fish_feel_pain.htm.

—— (2002b), 'The Neurobehavioral Nature of Fishes and the Question of Awareness and Pain', *Reviews in Fisheries Science*, 10: 1–38.

—— (2007), 'Anthropomorphism and ''Mental Welfare'' of Fishes', *Diseases of Aquatic Organisms*, 75: 139–154.

Searle, J. D. (1992), *The Rediscovery of the Mind*. Cambridge, MA; London: MIT Press.

—— (1994), 'Animal Minds', *Midwest Studies in Philosophy*, 19: 206–219.

—— (2007), 'Putting Consciousness back in the Brain', in *Neuroscience and Philosophy; Brain, Mind and Language*. New York: Columbia University Press.

Smith, A. D. (2005), 'Natural Kind Terms: A Neo-Lockean Theory', *European Journal of Philosophy*, 13: 70–88.

Sneddon, L. U. (2003), 'The Evidence for Pain in Fish: The Use of Morphine as an Analgesic', *Applied Animal Behaviour Science*, 83: 153–162.

Sneddon, L. U., Braithwaite, V. A. and Gentle, M. J. (2003), 'Do Fish have Nociceptors? Evidence for the Evolution of a Vertebrate Sensory System', *Proceedings of the Royal Society*, 270: 1115–1121.

Topál, J. and Csányi, V. (1999), 'Interactive Learning in the Paradise Fish (Macropodus Opercularis): An Ethnological Interpretation of the Second-Order Conditioning Paradigm', *Animal Cognition*, 2: 197–206.

Yue, S., Duncan, I. J. H. and Moccia, R. D. (2008), 'Investigating Fear in Rainbow Trout (Oncorhynchus Mykiss) Using the Conditioned-Suppression Paradigm', *Journal of Applied Animal Welfare Science*, 11: 14–27.

Wittgenstein, L. (1953/2000), *Philosophical Investigations*, trans. G.E.M. Anscombe. Oxford: Blackwell.

Zemach, E. (1976), 'Putnam's Theory on the Reference of Substance Terms', *Journal of Philosophy*, 73: 116–127.