17

Putnam's model-theoretic argument against metaphysical realism

BOB HALE AND CRISPIN WRIGHT

Metaphysical realism, as Hilary Putnam conceives it, is not a single, monolithic doctrine, but an amalgam of several closely associated philosophical ideas about the relations between language and reality, and between truth and knowledge or justifiable belief. One component on which Putnam places considerable emphasis is that even an ideal theory (a theory that is '*epistemically* ideal *for humans*' – ideal by the lights of the operational criteria by which we assess the merit of theories) may nevertheless be, in reality, false.[1] But commonly, Putnam presents metaphysical realism as involving adherence to three other claims, of which he takes this feature to be a consequence: that 'the world consists of a fixed totality of mind-independent objects', that 'there is exactly one true description of the way the world is' and that 'truth involves some sort of correspondence between words or thought-signs and external things and sets of things'.[2]

The so-called model-theoretic argument has played a leading role in the campaign Putnam has waged, in writings since 1976, against this outlook. Our leading questions will be: What is the argument? How is it best conceived as working? *Does* it work? Section I takes up the first, and gives our reasons for concentrating, thereafter, on the version of Putnam's argument set forth in his *Reason, Truth and History*. In Section II we explain how, in general terms, that argument is best conceived as working. Cursory inspection of Putnam's overall dialectic reveals it to incorporate three sub-arguments, collectively designed to show that the metaphysical realist confronts an insuperable problem over explaining how our words may possess determinate reference. In our next three sections we expound these three sub-arguments in more detail, and offer some critical reflections on them. Section III considers Putnam's version of the Permutation Argument, aimed at showing that reference cannot be determined by fixing the truth-conditions of whole sentences. In Section IV we then review his argument that reference cannot be fixed by our intentions or anything else 'in the head'; and in Section V we review his 'just more theory' argument, designed to show that the metaphysical realist cannot rescue the situation by appeal to causal or other natural connections between our words and the world. Having argued that the last of these arguments fails, we consider in Section VI whether Putnam's dialectical purposes might be better served by other, more specific arguments he has advanced elsewhere, aimed at showing that the project of giving a naturalistic account of reference is hopeless. In

Section VII we consider how the considerations adduced by Putnam might be seen as an argument telling selectively against metaphysical realism; and we conclude, in Section VIII, with a brief assessment of how far Putnam's argument, so viewed, may be taken to succeed.

I

There are significant differences between the versions of Putnam's argument in 'Models and Reality' (1977) and in *Reason, Truth and History* (1981a). Both confront the metaphysical realist with the same challenge – to show how words can stand in the determinate referential relations which his world-view demands. But the latter furnishes the more complete case for thinking the metaphysical realist incapable of meeting it. 'Models and Reality' deploys the Löwenheim–Skolem theorems, and closely related completeness results, to show – if all goes to plan – that no assignment of truth-values (however tightly constrained) to any (however comprehensive) class of whole sentences can suffice to fix the reference of terms and predicates. But there remain, so far as the argument of 'Models and Reality' goes, various ways a metaphysical realist may respond: for example, that speakers' intentions or other intentional states play an essential role; or that, if the reference of words is to be thought of as determined via their role in complete sentences, it is not those sentences' truth-*values*, but their truth-*conditions*, that matter. Further argument, of precisely the kind attempted in *Reason, Truth and History*, is needed to close off such moves.

In broadest outline, Putnam's thought in *Reason, Truth and History* has the following structure: If the world is to be conceived as consisting of 'some fixed totality of mind-independent objects', with truth consisting in 'some sort of correspondence relation between words or thought-signs and external things and sets of things',[3] then there must be determinate referential relations between the words and the things. But if so, the metaphysical realist owes an account of how that can be so. Putnam argues, by reviewing three, putatively exhaustive directions in which it might be sought, that there can be no such satisfactory account:

First, '*what goes on in the head' cannot determine what we are referring to.* We can imagine, Putnam suggests, a planet – Twin Earth – very much like Earth, populated by creatures very like ourselves, in surroundings very much like our own. There is, however, an interesting difference – the substance that fills Twin Earth streams and rivers, lakes and puddles, and comes out of Twin Earth taps, and so forth, is not H_2O but has a different chemical composition, XYZ. However, XYZ has just the same phenomenological properties as our water – it looks and tastes the same, and so on, and is, indeed, called 'water' on Twin Earth. If a Twin Earth dweller were somehow transported to Earth, she would not be able to tell our water apart from the liquid she encounters in similar circumstances back on Twin Earth. Her watery thoughts and experiences are, subjectively or 'from the inside', just like ours. In point of *pure* mental states relevant to the use of 'water' – that is, mental states identified neutrally with respect to the existence and character of such external things as might ordinarily get mentioned in their description – Twin Earth dwellers are indistin-

guishable from us. Yet when they speak of 'water', they are referring to XYZ, whereas we are referring to H_2O. So 'what's in the head' – pure mental states – does not determine reference. Reference varies in a way that cannot be explained by appeal to pure mental states. But to appeal to *impure* – world-involving – states would be just circular.[4]

Second, sub-sentential reference cannot be determined by fixing, via 'operational and theoretical constraints,'[5] either the truth-values *or even the truth-conditions* of whole sentences. This stronger conclusion is now obtained using more modest model-theoretic resources than in 'Models and Reality'. Given one scheme of reference which induces, at each possible world, such-and-such truth-values on complete sentences, we can obtain, by permutation, as many rival schemes as you like, which agree with the 'intended' scheme on the truth-values of whole sentences in *each* world, but diverge over assignments to terms and predicates.

Third, it is no use appealing to any further non-intentional – e.g. causal – condition as the needed source of referential determinacy. Any such appeal must assume, for example, that it is at least determinate what worldly relation our word 'causes' stands for. Saying that we use 'cats' to speak of just those things that stand in such-and-such causal relations to our use of the word is 'just more theory' – and, as such, just as liable to unwanted interpretations as anything else we may say.

II

There has been a tendency for commentators to interpret this train of thought as leading to a sceptical paradox comparable to that developed by Kripke in Wittgenstein's name (see Chapter 15, RULE-FOLLOWING, OBJECTIVITY AND MEANING, section 2): as Kripke's sceptic argues that there are no facts about meaning, so 'Putnam's Paradox' would have it that there are no facts about reference, all candidates for the constitution of such facts – the truth-conditions of sentences, speakers' intentional states, and causal and other forms of natural relationships between words and the world – failing to deliver the appropriately determinate goods. It is consistent with such an interpretation of Putnam's argument that he should think, as he certainly does, that the paradox admits of resolution, much as Kripke holds that there can be a solution to *his* Wittgensteinian paradox. But then the suggested parallel begins to limp. For one thing, it enjoins, taken strictly, that any solution will leave in place its sceptical conclusion – that there are no facts about reference – just as Kripke's sceptical solution leaves in place his sceptical conclusion, that there are no facts about meaning. If this were the intended form of Putnam's message, we should expect to find him explaining why/how it is that his preferred *internal realism*[6] can accept such indeterminacy with equanimity. But that is not what he does. What we find is, rather, the claim that the internal realist has no trouble *discounting* unintended interpretations of the sort that plague metaphysical realism.[7] Moreover, while it is true that metaphysical realism requires determinacy of reference – since without it, there appears to be no making sense of the claim that an ideal theory may yet be false – it appears that an outright demonstration of indeterminacy could not tell *selectively* against metaphysical realism. For while internal realism stops

short of claiming that even an ideal theory may be false, it will surely grant that a *less than ideal*, but still consistent, theory may be so. And this seems to require setting aside unintended interpretations just as much as does the metaphysical realist's more ambitious claim.

So what is the intended structure of the argument? It might be supposed that Putnam's purpose is not to explode the notion of reference altogether, but by engineering a *conditional* explosion – by showing that some distinctively metaphysical-realist assumption subserves a proof of indeterminacy – selectively to dispossess the metaphysical realist of the notion. The fact that, as will emerge, no specifically metaphysical-realist assumption oils the wheels of any of the three sub-arguments tells against this line: how, if so, could their combination spell trouble for metaphysical realism but leave internal realism unscathed?[8]

No: the right way to receive Putnam's argument, or so we suggest, is as turning on the crucial claim that *the metaphysical realist distinctively owes an explanatory/constitutive account of reference, but cannot deliver*. Much of what Putnam writes in *Reason, Truth and History* seems to confirm that this is indeed the primarily intended line of attack. The problem about reference, to which chapter 2 in the book is devoted, is repeatedly described as the problem of accounting for *how* the reference of our terms is fixed.[9] The emphasis throughout is on the need for explanation and the metaphysical realist's inability to supply one. Putnam writes:

> Of course the externalist agrees that the extension of 'rabbit' is the set of rabbits . . . But he does not regard such statements as telling us what reference *is*. For him finding out what reference *is*, i.e. what the *nature* of the 'correspondence' between words and things is, is a pressing problem . . . For me there is little to say about what reference is within a conceptual scheme other than these tautologies.[10]

The prevailing thought, then, would seem to be that the metaphysical realist incurs certain explanatory obligations which, for the internal realist, *simply do not arise* – that the internal realist may reasonably stay silent when questions are put about the *constitution* of the reference relation, about what makes it the case that a particular expression has the reference it does.

Why this should be so is a matter to which we shall return. But this, we shall assume, is how the overall gist of the argument should be interpreted.

III

Deferring issues about overall strategy, we now, in this section, review some of the detail of, and air some qualms concerning, perhaps the most arresting of the three ingredient claims in the *Reason, Truth and History* argument: the claim that even the *truth-conditions* of whole sentences containing them are insufficient to determine the references of sub-sentential expressions.

The well-known reasoning in support of this claim affirms that given any domain of objects, and a language used to speak about them, the references/extensions of the sub-sentential expressions of that language may be permuted

consistently with invariance in the truth-value assigned at each possible world to – hence in the truth-condition of – each sentence in the language. This is, Putnam suggests – though the claim needs discussion[11] – a generalization of Quine's contention in *Word and Object* that reference is inscrutable, based on the so-called 'Argument from Below' (see section 3 of Chapter 16, THE INDETERMINACY OF TRANSLATION). However, whereas Quine merely made a suggestive case that, for all our use of whole sentences containing it dictates to the contrary, 'rabbit' might refer to undetached rabbit parts, or temporal stages of rabbits, or the universal rabbithood – thus posing, at most, an unanswered challenge – Putnam's argument is wholly general ('rabbit' could, without change in the truth-conditions of any sentence containing it, refer to anything whatever) and, if correct, conclusive.

Let us review the illustration Putnam himself gives of the kind of thing that could be involved in such a systematic reinterpretation. Divide all possible worlds into just three kinds:

(a) worlds in which some cat is on some mat and some cherry is on some tree ('is on' is here tenseless)
(b) worlds in which some cat is on some mat, but no cherry is on any tree
(c) all other worlds.

Now fix the reference of 'cat' in a way which depends on which of these three groups the actual world belongs to. If the actual world is a type (a) world, then 'cat' is to refer to cherries and 'mat' is to refer to trees. If on the other hand the actual world is a type (b) world, then 'cat' is to refer to cats, and 'mat' is to refer to mats. Finally, if the actual world is a type (c) world, then 'cat' is to refer to cherries, and 'mat' is to refer to quarks.[12]

Now reflect that, if the actual world is as a matter of fact an (a)-world, in which some cherry is on some tree, the sentence, 'A cat is on a mat' will be true when the references of 'cat' and 'mat' are so stipulated. It will likewise be true in any (b)-world, since those are worlds in which some cat is on some mat, and, in those worlds, 'cat' and 'mat' have their customary reference. Finally, in (c)-worlds, the sentence will be false, since no cherry is on a quark. But these valuations, note, coincide exactly with those of 'A cat is on a mat' as ordinarily understood, with 'cat' and 'mat' assigned their customary reference. In short: the sentence 'A cat is on a mat' could have exactly the truth-conditions it does even if, for some possible worlds, including the actual world, 'cat' were to refer not to cats, but to cherries – or whatever you like.

Putnam shows[13] that this type of manoeuvre can be complicated so as to embrace simultaneously all the sentences of an entire language. And if sub-sentential reference may be varied in a systematic way without shift in truth-conditions, then whatever – if anything – determines reference, it cannot be the truth-conditions of whole sentences.

This conclusion is apt to seem deeply counter-intuitive. After all, are not the semantics of sub-sentential expressions exhausted by their contribution to the meanings of sentences containing them? So does not the reference of a term, or common noun, say, *have* somehow to be distinctively reflected in the meanings of

sentences in which it occurs? The argument gives pause, to say the least. We shall review four broad lines of reservation about it.

One quite common reaction is that Putnam's argument is somehow self-defeating. For in order to receive it as showing the existence of alternative interpretations of a language under which all its sentences retain their truth-conditions, we need already to be able to *grasp* the distinctions, generated by permutation, between the various interpretations. But if we can do that – if we can grasp and distinguish from one another the divergent interpretations on offer – then why can't we just *stipulate* that one among them in particular is the correct one? And why won't that stipulation be sufficient to render reference fully determinate? If, on the other hand, we can't make the requisite distinctions, then we are in no position to follow the reasoning by which Putnam seeks to persuade us that there is a difficulty.

It would be no answer to this to suggest that assumptions of determinacy of reference feature in Putnam's argument only for the purposes of *reductio as absurdum*. They don't. The claim that sentences' truth-conditions underdetermine sub-sentential reference, if supported by showing how particular permutation-based reinterpretations leave truth-conditions invariant, must depend, for its cogency, on a *continuing* grasp of the differences between the assignments of reference respectively involved in the various interpretations – a grasp which is to survive the drawing of Putnam's conclusion, and on which the grounds for that conclusion depend. So the thought may continue to seem impressive: if we understand the differences, then we can stipulate which interpretation is intended.

The main thing wrong with this objection is, rather, that it misconceives the point of Putnam's argument. It assumes that the argument, like Quine's, is properly seen as a *sceptical* one, directed against the determinacy of reference. Now of course, if that were its project, then the argument had better not proceed in a way which effectively presupposes determinacy, or employs materials which can be straightforwardly exploited so as to ensure it. But that is not Putnam's aim. Putnam's aim is to show, rather, that accounting for the determinacy of reference is a problem specifically for a particular kind of philosophical view. Accordingly – just so long as his own position is not vulnerable to the same difficulty – there is no reason why he should not argue in a way that presupposes determinacy. The intended gist of the permutation argument is merely that whatever secures a determinate reference for a particular sub-sentential expression, it is not the truth-conditions of the sentences in which it features. Putnam's position has to be that if any assumption of determinacy of reference is needed by the argument, it is an assumption which will eventually prove quite innocent from an internal-realist point of view. Of course, it's a good question why or whether that is so, and one to which we shall return.[14]

The second reservation is one some critics misguidedly advanced about the argument of 'Models and Reality'; that is, that the model-theoretic results to which that essay appeals are applicable only to first-order languages.[15] Strictly, this is so. It suffices to remark, however, that no such concern about generality affects the permutation argument of *Reason, Truth and History*. Given a permutation of the referents of the terms and compensating reinterpretation of the predicates of a language which preserves the truth-conditions – that is, the truth-values assigned

at each possible world – of each of its *atomic* statements, it is obvious enough that the truth-conditions not merely of first-order quantifications of those statements but also of their second-order generalizations, and indeed modalizations, will be likewise preserved.[16]

Perhaps more surprising is that the result will also extend to languages containing *intentional* operators, in particular expressions of propositional attitude. One might think that there would be a difficulty here, and that the scope of Putnam's argument would consequently have to be restricted. But this is not so. Take the hardest case: suppose that belief, for instance, is treated as a relation between a thinker and a *proposition*, and that any interpretation is required to assign as referent to a that-clause precisely that proposition which, in view of the assignments that interpretation makes to its subsentential parts, the clause in question comes to express. Thus, in Putnam's' illustration, 'that a cat is on a mat' comes to refer, in (a)-worlds, to the proposition that a cherry is on a tree and, in (c)-worlds, to the proposition that a cherry is on a quark, whilst keeping its usual reference otherwise (that is, in (b)-worlds). How might we set about gerrymandering an extension for 'X believes that' in order to ensure that 'X believes that a cat is on a mat' retains its actual truth-conditions – is true at just the worlds at which it is actually true – while the referents of 'cat', 'mat' and 'that a cat is on a mat' vary in accordance with the permutation in (this extension of) Putnam's illustration?

What is required, naturally, is that 'X believes that a cat is on a mat' should express a truth in all and only worlds in which X believes that a cat is on a mat. Now, since both cases are logically possible, there will be some (a)-worlds in which X does believe that a cat is on a mat and some in which he does not. We require accordingly that X stands, in just the former, in whatever relation the perverse interpretation assigns to 'believes that' to the proposition that a cherry is on a tree, and fails so to stand in just the latter. Clearly, therefore, we cannot leave the interpretation of 'believes that' invariant. For there have to be (a)-worlds in which X does believe that a cat is on a mat but does not believe that a cherry is on a tree, and in such worlds the truth-value of 'X believes that a cat is on a mat' would accordingly change under the permutation. Hence our reinterpretation will have to assign a new relation to 'believes that'. But what relation? Whatever the relation is, it will have to have the feature that *of necessity* a subject stands in it to the proposition that a cherry is on a tree just when he believes that a cat is on a mat. For if this is not a matter of necessity, then again, there will have to be (a)-worlds in which X does believe that a cat is on a mat, but does not stand in the relation in question to the proposition that a cherry is on a tree; and once again the perverse interpretation will get the truth-value of 'X believes that a cat is on a mat' wrong.

One's first thought is that there may simply be no such relation. But *that is not right*. There is, after all, at least the 'Cambridge' relation, in which a subject stands to the proposition that a cherry is on a tree just in case he believes that a cat is on a mat! One might compare this to the relation in which you stand to Mount Rushmore just in case you have seen a photograph of Snowdon. The crucial point is that permutation-based interpretation works *purely extensionally*. From the extensional

viewpoint, the latter relation has been fully specified just when what its extension is has been determined; and we have done that. For all and only our readers who have seen a photograph of Snowdon, the extension is the set of pairs, ⟨You, dear such reader; Mount Rushmore⟩. Similarly, the new relation which our perverse interpretation assigns to 'believes' will be one which has, at each (a)-world, an extension including the pair ⟨X; the proposition that a cherry is on a tree⟩ if and only if that (a)-world is one at which X believes that a cat is on a mat. For (b)-worlds – where 'that a cat is on a mat' is assigned as its referent the proposition that a cat is on a mat – no adjustment in the extension of the usual belief-relation is needed (at least, not in respect of X and the proposition that a cat is on a mat). Finally, at each (c)-world – where 'that a cat is on a mat' is assigned as its referent the proposition that a cherry is on a quark – 'believes' will be assigned an extension which includes ⟨X; the proposition that a cherry is on a quark⟩ if and only if that (c)-world is one at which X believes that a cat is on a mat.

This generality in the scope of the permutation argument is very striking. Arguably, however, the main thing one should conclude from it is how little the kind of 'interpretation' here in play has to do with *real* interpretation, as it were – interpretation in any sense which involves the specification of propositional contents which a thinker might conceivably have in mind. This is, in effect, the area of concern of a different line of objection – what we shall call the 'dilute truth-conditions' objection – to which we now turn. The objection concerns the ability of the permutation argument, even if this is sound as far as it goes, to deliver a conclusion of the intended significance. Since the goal of the argument, or so it may seem, ought to be to show that the reference of a sub-sentential expression is underdetermined by any features of the *meaning* of whole sentences containing it, Putnam must implicitly take it that he can encapsulate any germane notion of the meaning of a sentence in that of its 'truth-conditions'. To be sure, talk of 'truth-conditions' is, indeed, a standard philosophical idiom for gesturing at sentences' content. But Putnam's argument, the objection claims, works with so dilute a notion of 'truth-conditions' that this connection is subverted. Putnam's notion requires no more of truth-conditional equivalents than coincidence in their truth-values in all possible worlds – *strict equivalence*, in the sense of C. I. Lewis – and strict equivalence is intuitively quite consistent with manifest differences in semantic structure and content. In particular, their strict equivalence is insufficient to ensure that a pair of sentences make the same contribution to the content of sentences which embed them. Merely to take a pair of strict equivalents which draw on different conceptual resources – say, 'A and B are parallel' and 'Anything perpendicular to A is parallel to something perpendicular to B' – suffices to open up the possibility of someone who knows one but not the other. It follows that the truth-conditions, hence the content, of – to stay with the particular example – 'X knows that A and B are parallel' and 'X knows that anything perpendicular to A is parallel to something perpendicular to B' also differ. If we assume that the content of those two sentences is determined compositionally, there is then no alternative but to view the semantic contributions, and hence the meanings, of 'A and B are parallel' and 'Anything perpendicular to A is parallel to something perpendicular to B' as likewise different. And if there can be more to the

meaning – specifically, the semantic contribution to larger, embedding contexts – of a sentence than whatever it shares with its strict equivalents, then the general thought that the references of sub-sentential expressions may be determined by the *meanings* – in that richer sense, whatever it is – of the sentences which feature them is quite passed over by an argument which shows merely that *truth-conditions*, in Putnam's Lewisian sense, don't determine sub-sentential reference.

How may Putnam reply to this? He had better not challenge the inadequacy of strict equivalence to capture certain finer-grained but still intuitive notions of sameness and difference of sentence-meaning. Rather, what he ought to query is the stated characterization of the goal of the permutation argument: we should take the goal of the argument, that is, as that of showing, not that the reference of sub-sentential expressions is underdetermined by *any* features of the semantics of whole sentences containing them, but that sub-sentential reference is underdetermined by any whole-sentence semantical features *which can be explained without prior reliance on specific relations of sub-sentential reference*. That the reference of sub-sentential expressions might yet be recoverable from certain finer-grained semantical properties of sentences containing them – finer-grained than can be captured by relations of strict equivalence and non-equivalence – is accordingly, Putnam may charge, in no way inconsistent with his goal. For what goes into the constitution of such finer-grained semantic properties of a sentence will be, broadly, its mode of composition and the semantics – including reference – of its sub-sentential ingredients. Indeed, it is unintelligible how sentences could have such finer-grained semantic features in the first place unless we simply take for granted a gamut of relations of reference between sub-sentential expressions and items in the world. In short: the reply should be that while there may indeed be finer-grained conceptions of sentence-meaning than Lewisian strict equivalence, and while the reference of a particular sub-sentential expression may be recoverable from the finer-grained semantics of sentences containing it, this is all back to front from the point of view of answering Putnam's challenge. That challenge is to explain wherein the determinacy of sub-sentential reference is *constituted*. It is therefore irrelevant to appeal to semantical features of whole sentences which themselves depend upon the reference of those sentences' constituents.

Now, there is a possible misgiving about this reply connected with the question, mentioned earlier, of the extent of the analogy between Putnam's argument and Quine's 'Argument from Below'. The points of analogy are that the conclusion of both Putnam's and Quine's arguments may be expressed in the same way, that we can hold fixed the truth-conditions of a sentence while varying the reference of semantic constituents within it; and in both cases such constancy of truth-conditions may be glossed as consisting in the fact that, no matter how the world actually happens to be, the sentence will retain – after reference-permutation or Quinean reinterpretation respectively – the same truth-value as that secured for it by the (presumed) actual reference of its semantic constituents. However, Putnam's illustration would also seem to point to a potentially important difference. The kind of reinterpretation illustrated by the cats-and-cherries example sustains continuity in truth-value only because it is required to be sensitive to *what is actually the case*:

for instance, 'a cat is on a mat' is true, under the illustrated reinterpretation, in both type (a)- and type (b)-worlds only because what it says is *constrained to vary* as a function of which, if either, of those types the actual world belongs to. By contrast, any of Quine's alternative translation schemes for 'gavagai' (see section 3 of Chapter 16, THE INDETERMINACY OF TRANSLATION) will construe what the sentence says *in a uniform manner*, no matter what the actual world is like. In short, you cannot tell, under Putnam's assignment of reference, what 'a cat is on a mat' says unless you know how relevant matters stand in the world. But no such knowledge is needed to know the impact on 'gavagai' of any particular one of Quine's schemes. What follows is that for Putnam, but not for Quine, an *additional* distance would seem to be opened between preservation of truth-conditions and preservation of content: precisely, 'a cat is on a mat' retains its actual truth-conditions under the illustrated permutation – that is, has the truth-value it would actually have no matter which of the three types of world the actual world belongs to – only because what it says is *made to change* depending on which type of world that is. And while it may be acceptable for the argument to ignore differences in truth-conditions which can only be specified by presupposing differences in – and hence determinacy of – sub-sentential reference, it is still vital that the notion of 'truth-conditions' which it employs be as strong as possible consistently with that limitation. Yet the notion of sameness of truth-conditions at work in the permutation argument would seem to have *even less connection* with sameness of meaning than strict equivalence.

This development of the dilute truth-conditions objection probably ought to be open to just the same counter as the original objection. Suppose the objection is right that there is a perfectly good sense in which the effect of a Putnamian – in contrast to a Quinean – reinterpretation will be to have the *content* of a sentence vary as a function of what is actually true. The crucial question, however, is whether this variation in content could be appreciated from a standpoint which takes nothing for granted about sub-sentential reference, but is apprised only of *independently appreciable* semantic properties of whole sentences. What can be known about the semantics of a sentence by someone who knows nothing about the reference of its constituents? Could such a subject know more than its Putnamian truth-conditions, in which possible worlds it would be true and in which false? If not, then the claimed disanalogy between Quine and Putnam would not matter; the permutation argument would still be working with the strongest relevant notion of truth-conditions. Now there are, of course, things other than its Putnamian truth-conditions which someone can know about the semantics of a sentence who does not yet know anything about the reference of its constituents. In particular, there are all the things that are allowed to be available as data for a radical interpreter. Thus it is open to someone who does not yet know the reference of the constituents of 'a cat is on a mat' to observe its *use*, and to note in particular what appear to be its conditions of warranted assent. What he will observe if 'cat' refers to cats and 'mat' to mats is that the circumstances which prompt assent will tend to be those in which some cat is on some mat in a fashion salient to the assentor. But then, as may seem obvious – indeed, the whole point of Putnam's trick – the same pattern will still be expectable if 'cat' and 'mat' are assigned reference as in his

illustration. For suppose that is so and you are asked to assent to or dissent from 'a cat is on a mat'. Isn't it still true that you have only to consider whether you have reason to believe that a cat is on a mat? For if you do, then in both cases – when a cherry is on a tree (in which case that is what the sentence will say) and when none is (in which case it will say that a cat is on a mat) – you will have reason to think that 'a cat is on a mat' is true. So, of course, your observable pattern of assent will be the same.

Prima facie, then, the additional dilution would not matter in any case. However the decisive point is that, as the proofs in the Appendix to this chapter make clear, its apparently additional dilution of the notion of truth-conditions is actually an artefact of a dispensable – and it has to be said, misleading – feature of Putnam's illustration. There is no need for a permutation-based reinterpretation to 'kink' the assignments of reference after the fashion of the cats-and-cherries example. To be clear about this, consider a specific domain of objects, D, and, for simplicity's sake, restrict attention to all possible worlds involving just those objects and no others. Suppose we have a language, L, fitted to talk about the elements of D and to ascribe a given range of simple properties of them. A permutation of such a domain is simply a one-to-one mapping of D onto itself in such a way that no object need be correlated with itself; and the reinterpretation of the terms and (1-place) predicates of L associated with such a permutation does no more than have each term of L refer to the object onto which the permutation takes its actual referent, and have each predicate of L take as its new extension the set whose members are exactly the objects onto which the permutation takes the objects in its actual extension. Clearly, no matter what the actual extension of a predicate may be, the actual referent of a term will be a member of it only if its correlate under the permutation is a member of the set assigned to that predicate under the permutation-based reinterpretation. Although certain complications have to be finessed to take account of variation on the domains associated with different possible worlds, and of more complex predicates, this simple train of thought captures the essence of the permutation argument. And it points directly to a *uniform* reinterpretation of each sentence 'Fa' of L which is guaranteed to preserve its truth-value in any possible world. Where *p* is the permutation in question, that reinterpretation will read along the lines of 'the *p*-correlate of a is a member of the set of *p*-correlates of Fs'.

We conclude that Putnam has the resources to handle the dilute truth-conditions objection. But there is a related and fundamental worry still outstanding. The immediate effect of the permutation argument is that truth-conditions in, as we have seen, a somewhat technical sense underdetermine sub-sentential reference. And this result, we have stressed, is not to Putnam's purpose unless it bears interpretation as showing that all aspects of the use of a sentence that might be observed without presupposition about the reference of its constituents underdetermine what that reference may be. Now Putnam himself repeatedly expresses his finding as being that reference is underdetermined by both observational *and theoretical* considerations.[17] That is a very strong claim. It is tantamount to claiming not merely that alternative assignments of sub-sentential reference are *consistent with* all possible uses of a sentence, but that there will be *nothing to choose between*

them even when one takes account of all constraints, beyond empirical adequacy, which condition the construction of semantic theory. This has manifestly not been shown. It hardly seems likely, for instance, that, when all theoretical constraints on interpretation have been reckoned with, there still will be nothing to choose between interpreting speakers as expressing thoughts of the form: object a is F, and interpreting them as expressing thoughts of the form: the *p*-correlate – for some permutation, *p*, of the domain – of a is a member of the set of *p*-correlates of Fs! (For more on relevant such wider interpretational constraints, see Chapter 16, THE INDETERMINACY OF TRANSLATION, section 5.) In order for the permutation argument to succeed in showing that *best* interpretation of the use of whole sentences always has a variety of schemes of sub-sentential reference to select from, we have to be shown how to find *alternative extensionally coincident thoughts* to correspond to such monstrosities about *p*-correlates and sets of *p*-correlates – alternatives which it is as plausible, in the light of all relevant theoretical constraints, to interpret a subject as expressing by 'Fa' as the simple thought that a is F; and we have to be shown how to do this in a systematic way, right across the language. In short, to make good the suggestion that whole-sentence use underdetermines sub-sentential reference, permutation-based reinterpretations have to be shown to be, by *all* relevant constraints, as good – or anyway to facilitate reinterpretations which are as good – as standard interpretations. The results about permutation, by themselves, are powerless to show that is so.[18]

IV

We turn now to Putnam's argument that the intentional states of speakers are insufficient to determine the reference of their words. The argument, as we saw, proceeds by dilemma: if intentional states are conceived as 'pure' – so that, for instance, speakers both on Earth and Twin Earth can express the very same belief by 'Water is wet' – then reference may vary even though intentional states remain the same. If, on the other hand, intentional states are taken to be impure, so that the content of the belief that 'Water is wet' will be a function of the actual environment of its holders, then beliefs are now individuated by the actual references of the terms that occur in their expression, and thus presuppose, rather than constitute, such facts.

One cause for concern is whether the considerations offered in support of the first horn of Putnam's pure/impure dilemma can be made to cohere with what, later in the argument, he will want to say about the insufficiency of the sort of naturalistic conception of reference to which some – Hartry Field is an actual case[19] – may be tempted in response to Putnam's overall argument. To appreciate, after all, how reference may vary across environments in which the pure mental states of subjects remain the same, one has to have some conception of how reference *functionally depends* on environmental factors. But if such a conception is in place, then won't it constitute at least the beginnings of an account, independently of any play with speakers' mental states or the truth-conditions of sentences, of what it is that does determine reference? – precisely the kind of account which, according to the

third stage of Putnam's argument, cannot be given. Unquestionably there is a fair interpretative question here. The externalism about content which the first horn of the dilemma employs is a long-term theme in Putnam's writing; yet the metaphysical realist is apparently to be denied access to this element in Putnam's own philosophy in his attempt to respond to Putnam's challenge. However, we are entitled to proceed without pursuing that question by the consideration that this part of Putnam's argument in any case has no need to proceed in terms of the pure/impure dilemma.[20] A much simpler reflection will suffice. In order for speakers' intentional states, of whatever sort, to serve to establish the references of linguistic expressions, it has to be the case that the objects assigned to those expressions as their referents are already given as *objects of thought*. It is only as thought about – as referred to in thought – that we can fix, or understand, what it is for a particular object to be the referent of a particular symbol. But the constitutive question being put to the metaphysical realist arises no less for thought than for language. The challenge is to give an explanation of what it is for our thoughts to be of certain objects, rather than others, in the first place. The fact is, accordingly, that there never was any real option of the kind which the pure/impure dilemma is supposed to address. Intentional states cannot constitute reference. That our intentional states already have reference is (an aspect of) the problem, not its solution.

V

If the first two stages of Putnam's argument were to succeed, then the situation would be that no satisfactory constitutive account of reference can proceed in terms of facts concerning our intentional states, or facts about the truth-values, or even truth-conditions, of complete sentences or thoughts. To a metaphysical realist who is also a materialist, however, such conclusions would likely be entirely congenial, merely serving to underline the need for a quite different account of reference in broadly naturalistic terms. That it would be quite mistaken to think that any such account could meet metaphysical realism's needs is the burden of the third component of Putnam's argument.

This comprises, in fact, several distinguishable lines of attack: some of them are directed specifically at the idea that reference can be fixed by causal connections, but others aspire to greater generality, purporting to establish that there can be no 'reductive' explanation of reference in naturalistic terms or, more generally still, that once it is allowed that neither intentional states nor truth-conditions can form the basis of an explanation of how reference can be determinate, it can be seen that nothing else can do so either. The concern of this section will be with the most general – and most notorious – such line of all.

Putnam writes:

> Suppose there is a possible naturalistic or physicalistic *definition* of reference, as Field contends. Suppose
>
> (1) *x refers to y* if and only if *x bears R to y*

> is true, where *R* is a relation definable in natural science vocabulary without using any semantical notions . . . If (1) is true and empirically verifiable, then (1) is a sentence which is itself true even on the theory that reference is fixed as far as (and *only* as far as) it is determined by operational *plus* theoretical constraints. . . .
>
> If reference is only determined by operational and theoretical constraints, however, then the reference of '*x bears R to y*' is itself indeterminate, and so knowing that (1) is true will not help.[21]

Knowing that all instances of (1) are true won't help, Putnam thinks, because, by the permutation argument, they will remain true – and, indeed, will have the same truth-values in all possible worlds – when 'R' is taken instead to stand for a quite different relation R*. In fact, there are as many such alternatives R* as there are permutations of the universe of discourse. So supposing reference to be R has no more explanatory merit than supposing it to be R*. Hence it is merely an illusion that a unique reference relation has been singled out.

This move – of holding any attempted naturalistic characterization of reference to be 'just more theory', hostage to permutative reinterpretation – is one that Putnam repeatedly makes in the closing stage of his various attacks on metaphysical realism.[22] If allowable, it is of course decisive – for it will be available against *any* specific constraint the metaphysical realist may propose, regardless of its precise content, just so long as the constraint is formulated in a language to which the permutation argument applies.[23] The obvious and crucial issue is: is the move fair, or foul?

Well: it is foul, for a reason first stressed by David Lewis.[24] There is a distinction to be made between, on the one hand, an interpretation's *modelling* a proposed constraint – *making a statement of the constraint come out true* – and on the other, the interpretation's *actually conforming to that constraint*. The 'just more theory' gambit seems simply to miss this crucial distinction, taking the former for the latter.

To elaborate a little: Let C be some proposed (naturalistic) constraint on reference generally, L a language, S a sentence of L expressing C and I an interpretation of L. Suppose that I does indeed induce the value *true* on S. It might seem that I must conform to C; for S expresses C, after all, so that if I makes S come out true, isn't that just the same thing as I's conforming to C? Well obviously, not at all: we have no right to assume that, *whatever interpretation* of L is in play, S will (still) express C. Suppose, schematically, that S has the form:

$$\forall x \forall y \forall z\,(\text{speaker}(x) \,\&\, \text{expression}(y) \,\&\, \text{object}(z)) \rightarrow (x \text{ refers to } z \text{ by } y \rightarrow R(x,y,z))$$

One thing that may vary under different interpretations of L is, naturally, the relation assigned to 'R'. We may not take S as expressing C *tout court* – some interpretations will have S expressing C, others won't. An interpretation J under which S fails to express C may still make S come out true. And if we are able to express C in some other language L*, with resources sufficient to discuss the semantics of L, we may be in position to state – and it can be true – that while S is a sentence of L true under J, J does not conform to C.

A supporter of Putnam might reply that this will be a situation we can recognize as obtaining, and to which we can give expression, if, *but only if*, we can fall back on some other language L*, the reference of whose expressions can be assumed to be (sufficiently) determinate – in particular, there will have to be a sentence of L* by means of which we can give determinate expression to the intended constraint C. And it is precisely at this point, it may be alleged, that the metaphysical realist runs into 'just more theory' trouble. For his predicament is that any language, L* no less than L, will raise just the same problem about determinacy of reference. He can't just assume a more inclusive but referentially determinate L* in which it may be asserted that whilst the sentence S does indeed come out true on a whole host of interpretations (of L), all but a few of these are ones under which S fails to express the proposed constraint, to which they, furthermore, fail to conform. And if he can't simply assume that he can convey this thought in words, he cannot assume that he can think it either.[25] The upshot is, the supporter may claim, that while there is indeed a distinction of the sort Lewis proposes, the metaphysical realist cannot avail himself of it in the situation which matters, when any metalanguage, no less than the language with which we are originally concerned, gives rise to just the same difficulty.

But if this is the best reply that can be made, Lewis is right to cry 'Foul!' Just consider the dialectical situation. The metaphysical realist – Field or Devitt, for example – takes up the challenge to say what constitutes determinate relations of reference, only to find that no sooner has he opened his mouth than Putnam gags him with the complaint that he has no right to assume any of his words to be determinate in reference. The resulting situation is therefore really no different from that generated by the boring and jejune variety of meaning-scepticism which challenges an opponent to explain how meaningful discourse is possible, but won't countenance attempted answers because to presume them meaningful is to beg the question against it. Obviously the metaphysical realist has to be presumed capable of contentful – so, determinately referential – speech if he is to respond to Putnam's challenge, or indeed to any challenge at all. The onus legitimately placed upon him is not to *demonstrate that* determinate reference is possible, but to provide a constitutive account which *explains how* determinate reference works. Accordingly, he is perfectly within his rights to assume, at least pro tem, a metalanguage in which a determinate account of the putative mechanics can in principle be given.

VI

If the 'just more theory' move is illicit, that need still be no very serious matter for the overall argument provided there are good independent reasons for doubting that any naturalistic reduction of reference can be provided. Putnam has assembled in different places a variety of more specific arguments to this conclusion, of – so it seems to us – somewhat differing levels of cogency. We shall briefly review two lines in particular.

The first occurs in 'Model-theory and the "Factuality" of Semantics' (1989). The form of naturalistic proposal Putnam there envisages is familiar from such natural

scientific identifications as those of water with H_2O or of heat with mean kinetic energy of molecules. By identifying heat with mean molecular energy of motion, we accomplish what seems to be the best available explanation of empirically attested correlations involving variations in the temperature and pressure of a mass of gas whose volume is kept constant and so forth, and take this to sanction the identification. Might it not be, likewise, that by identifying the relation of reference with a certain physical relation, R, holding between tokenings of expressions and the worldly items to which they refer, we may achieve the best explanation of certain aspects of our use of those expressions? That is, why should ordinary scientific methodology not turn out to provide the same kind of case for identification of reference with R as for the identification of water with H_2O, or heat with mean molecular motion?[26]

Putnam's objection[27] is that any such proposal, grounded upon explanatory virtue, is viciously circular. Here is a key passage:

> One difficulty . . . is that this [proposal] uses the notion of *truth*. Our problem . . . was to explain how a particular reference relation – and that means, also, a particular extension for the notion of truth – gets attached to our words. To say that what does the attaching is the fact that certain sentences . . . are *true*, . . . is flagrantly circular. The problem, of course, is that what the semantic physicalist is trying to do is reduce intentional notions to physicalist ones, and this program requires that he not employ any intentional notions in the reduction. But *explanation* is a flagrantly intentional notion.[28]

We can discount what may seem to be the principal complaint in this passage. The general shape of the type of proposal mooted is that it is because a certain physicalistically specifiable relation R holds between our words and their referents that those words do in fact have those referents. It is, therefore, simply a misrepresentation to treat the proposal as asserting that the fact that certain *sentences* are *true* is what explains why our words refer as they do. That is, it seems quite gratuitous to impute to the physicalist the contention that what 'does the attaching' is the fact that a certain sentence (saying that our words bear R to some object) is true, rather than (simply) the fact which that sentence purports to state. This indifference to the distinction between object- and metalinguistic claims merely invites repetition of the main complaint already levelled at the 'just more theory' move.

The point about explanation made in the second half of the passage may seem more telling: if the semantic physicalist is in the business of giving a reductive account of reference in particular, and intentional notions in general, how can it be permissible to deploy intentional notions in so doing? But this too seems to us of doubtful force. Maybe the question would be appropriate if what was at issue was the standard type of *analytic* or *conceptual* reduction, a purported analysis of necessary and sufficient conditions of application. But the mooted form of proposal actually seeks an a posteriori reduction. It is anything but clear that all use of the notion of explanation must be eschewed, if the aim is that of saying what naturalistic relation between words and things in fact underpins reference – if what's on offer is a *theoretical* identification of reference with R, of the same general character as the

identification of heat with mean molecular motion.[29] All it seems to be legitimate to impose, by way of a general constraint, is rather that if use is made of an intentional notion in a statement which is part of a programme of physicalistic reductions of intentional notions generally, that use must be of a kind ultimately amenable to that form of reduction. It would be necessary to look at the details to see whether a particular physicalistic construal of reference, or explanation, violated this rather vague constraint. In any case, the constraints on legitimate a posteriori identification of properties and relations still remain to be clearly worked out.

Putnam is on much stronger ground, however, if it may be assumed that the naturalist proposal must ultimately identify reference with some *specific form of causal relationship* between the item or items that stand as the reference of a term and token uses of that term. Putnam himself, of course, has been prominent among those who have emphasized, as against the once orthodox Fregean conception of the matter, the role of causality in the determination of reference in a wide class of cases. But he warns us that neither his own proposal, nor Kripke's similar idea, were intended to explain from a standing start, as it were, how determinate reference is constituted; clearly they could not do so, since both pictures simply assume from the outset that individuals can be 'singled out for the purpose of a "naming ceremony"', and say nothing about how that might be done *ab initio*.[30]

Putnam has expressed various doubts about the viability of a reductive causal theory of reference. As he stresses, it will normally be the case that very many of the objects and events that figure in the causal ancestry of a particular utterance of an expression will not be what it refers to. Further, a term's or predicate's reference may be to, or may include, things with which it is not causally linked – items existing only in the future, for instance, are presumably available to be referred to but as yet sustain no causal relations.[31] Part of the problem for the causal theorist, then, is to single out the right causal relationship. Putnam is sceptical that this can be done in purely naturalistic terms without falling back on intentional notions. As against Evans's version of a causal theory, for example, according to which, roughly, a term refers to the dominant source of our beliefs involving it, he justly observes that the dominant source of our beliefs about electrons, say, may well be physics textbooks, rather than electrons themselves.[32]

Obviously these considerations are not conclusive. To the difficulty about future things, for example, it may be replied that in cases where the term introduced is general (perhaps a natural-kind term) it is to be understood that its extension comprises the causally connected samples and all other things of the same kind.[33] In general, causal theorists will surely agree that work is needed to characterize the appropriate kind of causal link – but why suppose the project to be hopeless?

Well, we suspect the project *is* hopeless. The core difficulty is to restrict, without ineliminable play with antecedent assumptions about its reference, the utterly disorderly mess of items that are apt to elicit tokenings of any given expression. In his Gifford Lectures[34] Putnam discusses probably the most sophisticated attempt to date to accomplish this: the proposal of Jerry Fodor[35] that the extension of a term comprises the smallest class of items which as a matter of natural law cause tokenings of the term, and whose doing so asymmetrically explains all other tokenings of

the term. For example, both horses and pictures of horses are apt to cause tokenings of 'horse'; but Fodor's intuition is that horses are the basic cause and therefore qualify as the reference, since it is only because horses cause tokenings of 'horse' that pictures of horses do.

Against this, it may be objected that there really is no clear priority as between 'If horses did not cause tokenings of "horse", neither would pictures of horses' and 'If pictures of horses did not cause tokenings of "horse", neither would horses'. Rather, what seems to be true is that it is because 'horse' refers to horses that *both* horses *and* pictures of horses – and thoughts of horses, and cows in a darkened field, etc., etc. – elicit, *ceteris paribus*, tokenings of 'horse'! In the jargon of possible worlds, the closest worlds in which pictures of horses do not cause tokenings of 'horse' are worlds in which horses don't either.[36] (For a fuller discussion of difficulties with semantic naturalism, see Chapter 5, 'A GUIDE TO NATURALIZING SEMANTICS'.)

VII

Let us try to take stock. First, to summarize the situation of the three sub-arguments of *Reason, Truth and History*. That the metaphysical realist has no option of explicating reference in terms of intentional states we take to be clear. However, the claim of the permutation argument to have shown that reference is underdetermined by features of the use of whole sentences is, as we saw, open to question. Moreover, the 'just more theory' move *is* a foul, and some of Putnam's own specific criticisms of causalist/naturalist proposals about reference are less than conclusive. However, to observe that the permutation argument as it stands is inconclusive for Putnam's purpose is one thing; but to make the kind of positive, constructive case for the determination of reference by whole-sentence semantics which, if such was her strategy, the metaphysical realist would need, is quite another. It is no clearer how such a case might in detail be made. Moreover, if that is not to be the strategy, then a causal account of reference – broadly construed – is the only remaining avenue to explore, yet the literature justifies nothing but pessimism about reconstructibility of semantic notions in non-intentional, causal terms. Putnam, then, may not have strictly proved all of his three lemmas. But he has done enough to issue a very pointed challenge, and one to which it is by no means clear that the metaphysical realist can satisfactorily respond.

Second, it merits emphasis that Putnam's considerations, even if conclusive, would provide no argument for the indeterminacy of reference as such: rather, what they would establish is that *if* referential relations had to be constituted in a certain kind of way – in the truth-conditions of sentences, for instance, or in causal connections – *then* reference would be indeterminate. The proper conclusion would be merely that a constitutive account of reference, of what makes it the case that a particular term, thought or spoken, stands for a particular object or kind, cannot proceed along any of the three lines reviewed. If those lines indeed exhaust the possibilities, then a case would have been made that there can be no fully explicit, reductive account at all of what constitutes the reference of a symbol to any particular item or range of items – at least, none which does not take for granted the

determinacy of reference of our thoughts as a background. 'Aboutness' would have to be conceived as primitive.[37]

Such a finding would no doubt be of great interest. But it will have been achieved, if it can be achieved, in a way that has no evident selective bearing on the status of metaphysical realism. The argument, if it can be made good, will be an argument for everybody. Moreover, notions which promise to admit of no reductive account are anyway ten a penny. So the questions remain: why see in the situation a (potential) *problem* to do with reference? And why, if so, a problem *distinctively* for *metaphysical realism?* The crucial task for a would-be sympathetic interpreter of Putnam is to provide convincing answers to these questions. How might such answers run?

To lack a constitutive account – and all prospect of a constitutive account – of a certain kind of subject matter is not, except in special circumstances, to have a reason to distrust its reality. That Putnam himself intends no scepticism about reference is abundantly clear from his willingness to allow that we can perfectly legitimately and fully adequately specify what 'cat', for instance, refers to: its reference is to *cats* (and therefore not to cherries, or to the *p*-correlates of cats under some permutation, *p*)! More generally, if it is granted that the language in which we are to state the reference of a term is an extension of the language to which that term belongs, then a homophonic formulation is a perfectly adequate response to someone who challenges us to individuate the reference of that term. If, however, that assumption is not granted – that is, if object- and metalanguage are distinct and the challenge is to justify the assignment of one scheme of reference to the terms of the object language, rather than to a permutation of it – then there are perfectly ordinary canons of interpretation to justify a preference, for example, for the assignment of cats to be the extension of 'chat' in French, rather than cats* – that is, cherries in a world in which cats are on mats and cherries are on trees. These will be canons which have to do, for instance, with the salience of cats in many of the situations which provoke 'chat'-talk among the French and a corresponding salient absence, for the most part, of cherries. That there are correct and incorrect things to say about what expressions refer to is enough for there to be *truths* – at least on the conception of truth favoured by the internal realist – about reference.

This is the key to the question of the selective bearing of the argument. What, precisely, might be put in doubt by the kinds of consideration reviewed is the existence of truths about reference *in a more substantial sense of* 'truth', a concept of truth whose applicability to claims of a certain kind requires, beyond the unimpeachability of those claims in the light of the ordinary discipline that informs their use, some form of robust *fit* between them and the world. For it is not enough for metaphysical realism merely that there be facts about what the expressions of our language refer to: these facts must be facts as metaphysical realism is wont to conceive *all* facts, facts no less sublime than – since constituted by relations to – the sublimated objects and properties which make up the metaphysical realist's world. There is accordingly no question of resting content with the sort of deflated account of them which is all that is provided by the homophonic platitudes and routine methodology of interpretation for which the internalist about truth may settle.

The metaphysical realist, then, owes a perspective on the nature of relations of reference which allows them to stand behind the routine interpretative methodology and which, indeed, explains its adequacy – explains how it is indeed a way of 'getting onto' or 'tracking' these independently constituted relations; a perspective which allows us to construe the truth of ascriptions of reference along robust correspondence lines, and which generally finds a place for such relations in the world as metaphysical realism conceives it. And there is, if Putnam's argument can succeed, no such perspective possible, because there is then nothing to be said about what reference *is*.[38]

In brief, then, we have a rich and complex argument to the conclusion that reference admits of no reductive account, coupled with the claim that metaphysical realism – but not internal realism – is saddled with a world-view that cannot be properly understood unless such an account can, *per impossibile*, be given. The crucial difference is entirely one of explanatory obligation. For metaphysical realism, reference is a matter of relations between robustly distinct existences, items of language and thought on the one side, and items in a stubbornly alien world on the other; and this conception, Putnam's driving idea has it, entrains a commitment to the possibility of some sort of external perspective on the nature and constitution of this relationship – exactly what, if his argument succeeds in detail, cannot be delivered. So the metaphysical realist must, in the end, be driven to obscurantism: a conviction in the reality of relations constituted, he knows not how, between his thought and a world wholly alien to it.[39]

VIII

Why does internal realism incur no parallel obligation? Can the mere currency of standards of correctness for claims about reference really ensure that no issue arises? It is one thing to get a sense of Putnam's thought on this point; another to determine whether it is really convincing. The key idea seems to be that, as Putnam repeatedly expresses it, 'there is no ready-made world': that the division of the world into particular objects and kinds of thing is somehow coeval with, rather than merely *reflected by*, the divisions among our concepts and the expressions for them. If the kind picked out by a term of ours is thought of as originally constituted quite independently of the use of that term and the conceptual resources associated with it, then the question has to arise: what attaches the term on to just that kind, as opposed to another? That is the question which metaphysical realism is charged to answer. If, by contrast, the kind is regarded as in some way having no being independently of our deployment of those very conceptual resources, then there is no real linkage to explain, any more than it wants an explanation, how the patterns on a slide manage to be congruent with the images it casts upon a blank screen.

This kind of simile is convincing enough in its way. The difficulty is to give it substance in the case that matters – to see what the idea that human conceptual activity 'slices up' the world really comes to.[40] But perhaps on reflection there is

room to repudiate the metaphysically realist conception of a 'hooking' of language onto a sortally predeterminate world without recourse, natural though it may be, to opposing constructivist metaphors.[41] The crucial point is that, unless the unity of a range of items is in some way fixed in advance of the institution of using a term of which they are the reference, there is no non-trivial question what makes for the connection between that term and that range: the range of items in question just constitutionally is that for which the term in question stands.

That leaves the metaphysical realist the options of faulting the detail of the stages of the argument, or living with its conclusion: that to conceive of the world in a certain kind of robustly autonomous fashion is to consign the relation between the vehicles of our thought and the taxonomy of the world to unaccountability. Putnam effectively ridicules such an upshot. But ridicule, it may be countered, is no substitute for argument. Any broad philosophical system will have its primitive notions and theses. Further argument may be demanded as to why metaphysical realism may not legitimately go primitive at the interface between language and the world. That is what it must do if intentionality – 'aboutness' – is indeed irreducible, as in effect the three ingredients of Putnam's argument combine (if they are sound) to show. To be sure, no aspirant to a purely physicalist version of metaphysical realism could rest content with primitively intentional relations of aboutness. And Putnam may be right to say that 'materialism is the only *metaphysical* picture that has contemporary "clout". Metaphysics, or the enterprise of describing the "furniture of the world" . . . has been rejected by many analytic philosophers . . . Today, apart from relics, it is only materialists (or "physicalists", as they like to call themselves) who continue the traditional enterprise'.[42] But it remains to be convincingly explained why 'the only sort of metaphysical realism that our time can take seriously'[43] should be a thorough-going physicalism, or why irreducible intentionality should be especially uncomfortable for one of metaphysically realist predilection.

We end with one final reservation about the scope of the argument. If the interpretation offered is sound, then it can engage only a realist who accepts the autonomy of the division of the world into objects and kinds. So far as we can see, it must therefore fail to touch an intermediate, apparently coherent combination of views: the combination which yokes rejection of the idea that there is a 'pre-sliced', 'ready-made world' – that the world divides into kinds of thing, stuff and so on quite independently of our efforts to devise a conceptual scheme in terms of which it may be best described and understood – with acceptance of an evidentially unconstrained conception of truth, that is, with realism in the sense Dummett has made familiar (see Chapter 12, REALISM AND ITS OPPOSITIONS). Putnam has sometimes written as if the latter form of realism must fall to his argument. If we are right, that can be so only if a Dumettian, realist conception of truth must in the end consist in the kind of robust correspondence conception which is the essence of metaphysical realism as Putnam conceives it. However, it is one thing to accept that questions about what words refer to make sense, and have determinate answers, only within a conceptual scheme (so that the words cannot be thought of as having reference to

an antecedently determinate world of objects and kinds), and another to claim that we cannot combine those words into statements which may, in principle, possess determinate but undetectable truth-values. If the latter *is* a consequence of the former, further argument is needed to show it.[44,45]

Appendix: permutation results

In his Appendix to *Reason, Truth and History*, Putnam shows how to prove a relatively strong permutation result to the effect that, given an interpretation I of a (first-order) language L, we can construct another ('unintended') interpretation J which preserves the truth-conditions of all the sentences of L (in his sense, under which sentences have the same truth-condition if they have the same truth-value at all possible worlds), whilst varying the extensions of terms and predicates. Here, we first prove a more basic, weaker result (to the effect that, given an interpretation of a first-order language, we can always construct an alternative 'unintended' interpretation which coincides with the given interpretation over the truth-values of all the sentences, while varying the extensions of terms and predicates). We then indicate how the method of proof (which differs somewhat from that employed by Putnam) may be extended to obtain, first, a result essentially the same as Putnam's and then some stronger results, for second-order languages and for languages with modal operators.

Weak permutation

For this, we work with a first-order language L, with logical constants: $\neg$, $\wedge$, $\exists$; terms, comprising individual constants a, b, c, . . . and variables x, y, z, . . . ; and predicate constants F, G, H, ... The atomic sentences are just the strings $Ft_1, \ldots, t_n$, consisting of an n-place F followed by n occurrences of individual terms. If A,B are sentences, so are $\neg A$, $A \wedge B$ and $\exists x A(x)$, where x is any variable and A(x) comes from some sentence A by replacing one or more occurrences of some one individual constant by occurrences of x.

An interpretation I of L consists of a non-empty domain D with assignments of elements of D as denotations of the individual terms and of sets of ordered n-tuples of elements of D, for appropriate choices of n, as extensions of the n-place predicates. Thus to each 1-place predicate, I assigns as extension a subset of D – intuitively, the set of elements of D having the property for which, under I, that predicate is take to stand; to each 2-place predicate, I assigns a set of ordered pairs of elements of D – intuitively, the pairs of elements of D the first of which bears to the second the relation for which, under I, that predicate stands, and so on. '$I(A) = 1$' denotes that A is true under I, and is defined as follows:

$I(Ft_1 \ldots t_n) = 1$	iff	$\langle I(t_1) \ldots I(t_n)\rangle \in I(F)$
$I(\neg B) = 1$	iff	$I(B) \neq 1$
$I(B \wedge C) = 1$	iff	$I(B) = I(C) = 1$
$I(\exists x B(x)) = 1$	iff	there is an interpretation I° which differs from I at most in its assignment to x, such that $I^\circ(B(x)) = 1$

Theorem 1 (weak permutation)

Let I be any interpretation with domain D, and ϕ be any permutation of D. Let I* be any interpretation with the same domain D such that, for every term t, $I^*(t) = \phi(I(t))$, and for every n-place F, $I^*(F) = \{\langle d_1, \ldots, dn\rangle \mid \langle \phi^{-1}(d_1), \ldots, \phi^{-1}(d_n)\rangle \in I(F)\}$. Then for any A, $I(A) = 1 \leftrightarrow I^*(A) = 1$

Strictly, for the purposes of Putnam-type arguments, we need only establish that for given I, D and ϕ, there is at least one interpretation meeting the specified conditions on I*, for which the theorem's consequent holds. However, the proof can proceed more smoothly for the theorem as stated. It is obvious that there *are* (non-trivial) interpretations meeting the antecedent conditions.

Proof is by induction on the degree of A, as measured by the number of logical operators occurring in it. So the induction hypothesis (IH) is that the theorem holds for all wffs of degree <A, and on this hypothesis it is to be proved that the theorem holds for A. More fully stated, IH is:

> If I^1 and I^2 are *any* interpretations with the same domain, such that for each term t, $I^2(t) = \phi(I^1(t))$ and for any n-place F, $I^2(F) = \{\langle d_1, \ldots, d_n\rangle \mid \langle \phi^{-1}(d_1), \ldots, \phi^{-1}(d_n)\rangle \in I^1(F)\}$, then for any B of degree <A, $I^2(B) = 1 \leftrightarrow I^1(B) = 1$

A is atomic i.e. $Ft_1 \ldots t_n$ for some n.

$I^*(Ft_1 \ldots t_n) = 1$ iff $\langle I^*(t_1) \ldots I^*(t_n)\rangle \in I^*(F)$
iff $\langle \phi(I(t_1)) \ldots \phi(I(t_n))\rangle \in I^*(F)$
iff $\langle \phi^{-1}(\phi(I(t_1))) \ldots \phi^{-1}(\phi(I(t_n)))\rangle \in I(F)$
iff $\langle I(t_1) \ldots I(t_n)\rangle \in I(F)$
iff $I(Ft_1 \ldots t_n) = 1$

Induction step for $\exists$

Suppose $I(\exists xB(x)) = 1$. Then for some I° differing from I in at most its assignment to x, $I^\circ(B(x)) = 1$. Let $I^\#$ be the same as I* except possibly over its assignment to x, where $I^\#(x) = \phi(I^\circ(x))$. It is easily verified that I° and $I^\#$ meet the conditions on I^1 and I^2 in the induction hypothesis, which then yields that $I^\#(B(x)) = 1$. Hence $I^*(\exists xB(x)) = 1$. The steps are obviously reversible.

Other cases for induction are straightforward. □

Theorem 1 ensures that given any assignment of truth-values to the sentences of L_n induced by an interpretation I, there will be a quite different interpretation I* of L based on a permutation of I's domain, which induces all the same truth-*values* on L's sentences, but makes quite different assignments to the names and predicates of the language.

Strong permutation

A stronger permutation result will be that given an interpretation I of L_n we can get a different interpretation I* that departs from I over its assignments to names and

predicates, whilst giving L's sentences the same truth-*conditions* (in the Putnam sense – the sentences of L coincide in truth-value not just at the actual world, but at every possible world, under the two interpretations). To state and prove this stronger result, we need some preliminary stage-setting:

By a world structure we mean a triple $\langle D,W,\sigma\rangle$, where D and W are non-empty sets (intuitively, think of W as the set of all possible worlds, of D as a very inclusive set of objects, containing each object which exists at any of the worlds in W), and σ is a function from W into the non-empty subsets of D (i.e. σ assigns a non-empty subset of objects to each world[46])

Interpretations I are now assignments as follows: for each i and j, I assigns to the term t_i an element of the domain of w_j as its denotation relative to that world, i.e. $I(t_i,w_j) \in \sigma(w_j)$. And to each n-place F, I assigns, relative to each world w_j, a set of ordered n-tuples from the domain of w_j, i.e. $I(F,w_j) \subseteq (\sigma(w_j))^n$.

Truth under I is now of course a relation between sentences of L and worlds, defined thus:

Atoms	$I(Ft_1 \ldots t_n, wj) = 1$ iff $\langle I(t_1,w_i) \ldots I(t_n,w_j)\rangle \in I(F,w_j)$
Molecules	$I(\neg B,w_j) = 1$ iff $I(B,w_j) \neq 1$
	$I(B \wedge C,w_j) = 1$ iff $I(B,w_j) = 1$ and $I(C,w_j) = 1$
	$I(\exists xB(x),w_j) = 1$ iff there is an interpretation I° which differs from I at most in its assignment to x, such that $I^\circ(B(x),w_j) = 1$

Theorem 2 (strong permutation)

Let I be an interpretation of L. Let the ϕ_i be permutations[47] respectively of each of $\sigma(w_i)$ for all the $w_i \in W$. Let I^* be any interpretation of L such that for all i and j, $I^*(t_i,w_j) = \phi_i[I(t_i,w_j)]$ and for every n-place F, $I^*(F,w_i) = \{\langle d_1 \ldots d_n\rangle \mid \langle \phi_i^{-1}(d_1) \ldots \phi_j^{-1}(d_n)\rangle \in I(F, w_j)\}$. Then $I^*(A,w_i) = 1 \leftrightarrow I(A,w_j) = 1$

Proof is again by induction on the degree of A – the foregoing proof of *weak permutation* is readily adapted to show what's required for arbitrary w_j, simply by writing in w_i as an extra parameter as appropriate.

A is atomic i.e. $Ft_1 \ldots t_n$ for some n:

$I^*(Ft_1 \ldots t_n,w_j) = 1$

iff $\langle I^*(t_1,w_j) \ldots I^*(t_n,w_j)\rangle \in I^*(F,w_j)$

iff $\langle \phi_j(I(t_1,w_j)) \ldots \phi_j(I(t_n,w_j))\rangle \in I^*(F,w_j)$

iff $\langle \phi_j^{-1}(\phi_j(I(t_1,w_j))) \ldots \phi_j^{-1}(\phi_j(I(t_n,w_j)))\rangle \in I(F,w_j)$

iff $\langle I(t_1,w_j) \ldots I(t_n,w_j)\rangle \in I(F,w_j)$

iff $I(Ft_1 \ldots t_n,w_j) = 1$

As before, the induction step is quite straightforward. Here, for illustration, is the case for $\wedge$:

Suppose $I(B \wedge C,w_i) = 1$. Then $I(B,w_i) = I(C,w_i) = 1$. By IH, $I^*(B,w_i) = I^*(C,w_i) = 1$. Hence $I^*(B \wedge C,w_i) = 1$. Steps obviously reversible.

Strengthening for second-order languages

We extend our first-order language L by permitting binding of (first-level) predicate variables by the second-order existential quantifier $\exists f$ – we use f,g, . . . as predicate variables. An interpretation of our second-order language L^2 will make assignments to them of entities of the same types as are assigned to predicate constants in the first-order case. 'true under I' is defined as for previous cases, except that we add a clause for the second-order quantifier:

$I(\exists fB(f)) = 1$ iff there is an interpretation I° which differs from I at most in its assignment to f, such that $I^\circ(B(f)) = 1$

With this addition, we can straightforwardly extend the weak and strong permutation results to the second-order case – all that is needed is an extra case in the induction, dealing with sentences in which the principal operator is second-order $\exists$. For the second-order extension of Theorem 1, this runs:

Induction step for second-order $\exists$

Suppose $I(\exists fB(f)) = 1$. Then for some I° differing from I in at most its assignment to f, $I^\circ(B(f)) = 1$. Let $I^\#$ be the same as I^* except possibly over its assignment to f, where $I^\#(f) = \{\langle d_1, \ldots, d_n\rangle \in D^n \mid \langle \phi^{-1}(d_1), \ldots, \phi^{-1}(d_n)\rangle \in I^\circ(f)\}$. Then by the induction hypothesis, $I^\#(B(f)) = 1$. Hence $I^*(\exists fB(f)) = 1$. The steps are obviously reversible.

Languages with modal operators

The addition of a modal operator, say $\Box$, to L (or L^2) permits the formation of complex sentences which are not truth-functions of their atomic constituents. That is, we can form sentences B with atomic consituents $A_1 \ldots A_k$ so that B's truth-value at a world w_i is not a function simply of the values of $A_1 \ldots A_k$ at w_i. B's truth-value at w_i is, rather, a function of the values of $A_1 \ldots A_k$ at the other worlds in W.

Does this prevent us from running the permutation argument? Well, it seems that it should *not* do so – just because, while a modal sentence's truth-value at a given world is not a function of the values of its atomic ingredients at *that* world, it *is* a function of their values at other worlds. But we know from *strong permutation* that we can jiggle the assignments to individual constants and predicates in such a way as to obtain an 'unintended' interpretation which agrees with the original interpretation on the truth-values of all the sentences of L (or L^2) at all possible worlds (so that they have the same truth-conditions, in Putnam's sense). It follows from this that adding $\Box$ to L (or L^2), with the usual clause to the effect that $I(\Box B,w_i) = 1$ iff $I(B,w_k) = 1$ for all w_k accessible from w_i, can make no essential difference to the situation. The essential point is this. Given an interpretation I which induces a

pattern of truth-values on a sentence B across the possible worlds, we can construct a variant interpretation I*, differing from I in its assignments to terms and predicates (and in case of L^2, predicate variables) at those worlds, but agreeing with I on the induced value of B at each world. And that is enough to ensure that I and I* will not diverge over the truth-values of modal functions of B.

Notes

1 'Models and Reality', in Putnam (1983, p. 13).

2 Putnam (1981a, p. 49); cf. also 'A defense of internal realism', in Putnam (1990a, p. 30).

3 Putnam (1981a, p. 49).

4 The Twin Earth argument was first presented in Putnam's 'The Meaning of "Meaning"', see especially Putnam (1975, p. 223) and following. An abbreviated statement of it is given in Putnam (1981a, pp. 22–9). See also pp. 41–3 for the distinction between pure and impure mental states: and Chapter 5, A GUIDE TO NATURALIZING SEMANTICS.

5 By saying that an assignment of truth-values to sentences meets operational constraints, Putnam means, roughly speaking, that it accords with all the observational data that is available in principle. By theoretical constraints he means whatever further methodological constraints – including pragmatic considerations such as simplicity and economy – guide the optimum choice between theories which meet all operational constraints. Cf. 'Models and Reality', in Putnam (1983, pp. 3–6).

6 See Putnam's classic characterization of the 'internalist perspective' (1981a, pp. 49 and following).

7 Thus he writes:

> For an internalist like myself, the situation is quite different. . . . signs do not intrinsically correspond to objects, independently of how those signs are employed . . . 'Objects' do not exist independently of conceptual schemes. *We* cut up the world into objects when we introduce one or another scheme of description. Since the objects *and* the signs are alike *internal* to the scheme of description, it is possible to say what matches what. . . . Indeed, it is trivial to say what any word refers to within the language the word belongs to, by using the word itself. What does 'rabbit' refer to? Why, to rabbits of course. (1981a, p. 52)

See also 'Models and reality', in Putnam (1983, p. 24).

8 This question has exercised some of Putnam's critics, e.g. Blackburn (1994, p. 27), but needlessly, if we are right.

9 See e.g. remarks at pp. 25, 27 and 29.

10 Putnam (1981a, p. 52). Cf. also:

> if the received view is correct, then we would have an elegant *account of how* intensions and extensions are fixed [p. 32, our emphasis] One might say that . . . my 'mental representations' . . . *refer* to cathood . . . this may be true, but it just repeats that reference is fixed in one way rather than another. This is what we want to explain and not the explanation sought. [p. 37] To explain reference in terms of (impure) intention would be circular. And the problem of how *pure* mental states of intending, believing, etc., can . . . constitute reference is just what we have found so puzzling. [p. 43]

11 More about this matter below.

12 Cf. Putnam (1981a, p. 34). Putnam's stipulation for (c)-worlds is a little odd – it would have sufficed to have 'cat' refer to cats and 'mat' refer to mats in this case, since all that's required is that 'A cat is on a mat' be false in (c)-worlds.

13 For formal details, see the Appendix.

14 That said, it's worth observing that, even if Putnam's project were to argue for the indeterminacy of reference *tout court*, it's not clear that the permutation argument would be vulnerable to the stated objection. For the proof of the permutability of reference – illustrations apart – is *entirely general*, and following it need involve consideration of no specific suppositions about the reference of particular expressions in the language: suppositions whose status might then be settled by stipulation. Someone – not Putnam – who wanted to harness the permutation argument to a general scepticism about reference *could* quite coherently carry its conclusion forward in the form of the counterfactual: if there were such a thing as determinate reference, it would not be recoverable from the truth conditions of sentences. And indeed, the overall strategy of arguing for indeterminacy by establishing enough such counterfactuals, with a sufficient variety of consequents ('. . . , it would not be recoverable from speakers' intentions', '. . . , it would not be recoverable from facts about causality', etc.), is a perfectly coherent one. By the same token, though, the concern – for a supporter of Putnam – that the model-theoretic argument may fail stably to focus against metaphysical realism, dissolving instead into 'Putnam's Paradox', is not so easily set aside.

15 See, for instance, Ian Hacking (1983, p. 105); though this may not be quite fair to Hacking, who in the relevant passage is mainly raising a doubt about the first-order formalizability e.g. of physical theory, and is not really emphasizing the failure of the Lowenheim-Skolem Theorem at second order. Cf. Putnam's remarks in note 11 of Putnam (1989, p. 230).

16 Permutation results for second-order languages and languages with the usual modal operators are outlined in the Appendix.

17 See e.g. Putnam (1989, p. 215).

18 There are, however, reasons to qualify the force of this reservation, whose significance will emerge only when more has been done to explain how Putnam's argument can bear selectively on metaphysical realism. See Section VII, and especially n. 38 below.

19 Hartry Field (1972). Field's view is discussed by Putnam (1981a, pp. 45–6; 1978, pp. 14–17, 30–32 and 57–58).

20 A reason for thinking the tension merely apparent will anyway emerge in Section VI below – see also note 30.

21 Putnam (1981a, pp. 45–6).

22 Besides directing it at Field's in the passage quoted, Putnam (1983, p. 18) makes essentially the same move against Evans's version of the causal theory, and (1989, pp. 219–20) against Devitt's appeal to a causal theory.

23 This claim appears to run counter to Putnam's own view, as expressed in 'Model Theory and the "Factuality" of Semantics' (1989). He stresses there that his model-theoretic argument is directed against a limited target – *physicalistic* metaphysical realism. Certainly some of the argumentation rehearsed in that paper relies upon the assumption that the metaphysical realist aspires to a physicalist account of reality – including the circularity argument discussed above. Our point is that the 'just more theory' move is *not* subject to this limitation. A similar point is made by David Lewis (1984, pp. 232–3).

24 Ibid., pp. 224–5.

25 Something like this may well be the intended thrust of Putnam's complaint (1983, p. xi) that the causal realist 'ignores his own epistemological position'.

26 Cf. Putnam (1989, pp. 216–17).

27 Putnam advances two quite distinct objections against the proposal. This is his first objection; we shall discuss the second in due course. Meanwhile, note that the first objection is to any identification of reference with a physicalistic relation, regardless of whether it is made in the interests of defending metaphysical realism.

28 (1989, p. 217). Putnam has 'requirement' where we have 'proposal'. The requirement to which he refers is presumably that if a relation R is to be the 'intended' reference relation, the supposition that R is the reference relation should yield an explanation of facts about our use of words.

29 A more detailed formulation of this argument is given in 'Beyond Historicism' (1983, pp. 290–98 and 292–5). We lack space to discuss it here, but it seems to us that it is vitiated by the same gratuitous assumption that anyone who proposes a 'theoretical identification' of an intentional notion – such as *explanation* or *reference* – is thereby debarred from using the notion in question in arguing for the identification. That this might be a reasonable restriction to impose on attempts at *analytic* reductions of intentional notions seems quite irrelevant.

30 Cf. 'Models and Reality' in Putnam (1983, p. 17). This bears on the interpretative question left dangling in Section IV – it is clearly quite consistent with holding that a causal constraint needs to be met in many (or even all) cases of genuine reference to deny that a full constitutive account of reference may be given in purely causal terms.

31 Here is a relevant passage from 'Model Theory and the "Factuality" of Semantics' at p. 219:

> if *E(T)* is the event of someone's using a token of a term *T*, then there is a good sense of 'causal connection' in which *every* event in the backward light-cone of *E(T)* is 'causally connected' to the event *E(T)*; but it will almost never be the case that the term *T* . . . *refers* to every event in the backward light-cone of *E(T)* (and it will typically be the case that the term does refer to things with which the token is *not* causally connected, e.g., future things).

32 'Models and Reality' in Putnam (1983, p. 18). See Evans (1973, pp. 187–208) for his version of the theory. For some further discussion of this approach, see Chapter 21, REFERENCE AND NECESSITY, esp. sect. 4.

33 Putnam would probably concede that the first difficulty may not be insuperable – cf. his acknowledgement that Evans has a proposal, to which he offers no objection – to deal with this problem; see the footnote on p. 18 of 'Models and Reality', in Putnam (1983). He does press the difficulty a little further, claiming that the distinction between causes and background conditions is inescapably interest-relative; but this shows at best that the relevant causal relations can't be singled out by appeal to that distinction, not that they can't be singled out at all.

34 Putnam (1992, ch. 3).

35 Fodor (1990).

36 Further objections, complementing those brought by Putnam (and those discussed in Chapter 5, A GUIDE TO NATURALIZING SEMANTICS), and including a forceful play with the holism of the mental, are developed by Paul Boghossian (1991).

37 Putnam, of course, is well aware of the possibility of this response to his argument, envisaging it explicitly (1989, p. 220); however, he does not regard acknowledging the primitiveness of reference as a commitment to regarding it as 'simple and

irreducible'. *Representation and Reality* is, in effect, an extended argument to the contrary.

38 The astute reader will note that if these considerations are indeed the key to the question of how Putnam's argument can tell selectively against metaphysical realism, then there actually is little force – in the resultant dialectical setting – in the reservation with which our discussion of the significance of the permutation argument concluded in Section III. That reservation was, in effect, that while permutation-based reinterpretations of a language might be *consistent* with all data concerning the use of its sentences, they would be likely to be dominated by the preferred interpretation once appropriate constraints on the construction of interpretational theories, beyond adequacy to the linguistic data, are allowed their proper influence. But that, if correct, is a point about the methodology of interpretation – something which can be freely acknowledged from the internal realist point of view as conditioning the concept of truth that applies to ascriptions of reference and other semantic claims, but which takes us no closer to the *constitutive* account which the metaphysical realist needs of the nature of reference, conceived as a network of external relations of which methodologically superior interpretation is, at best, a means of discovery.

39 Recall the complaint which Putnam airs against Lewis's positive view, that it amounts to 'saying that we-know-not-what fixes the reference relation we-know-not-how' (1989, p. 220).

40 So far as we are aware, Putnam does not himself explicitly employ the metaphor of 'slicing'. But it is common in discussion of his ideas and implicit in several of his own characterizations of internal realism. For example, '"Objects" do not exist independently of conceptual schemes. *We* cut up the world into objects when we introduce one or another scheme of description.' (Putnam 1981a, p. 52.) Of a piece with this are his frequent characterizations of external or metaphysical realism as involving – via its commitment to the idea that there is, whether we can discover it or not, just one true theory of the world – a belief that there is a 'ready-made world', having an intrinsic or 'built-in' structure, comprising a 'fixed totality of mind-independent objects'. Cf., for example, Putnam (1983, p. 211, and 1981a, p. 49).

41 But an outright repudiation of the idea of sortal predetermination, even if not accompanied by a lurch into constructivist metaphor, would be in at least prima facie conflict with retention of the idea, of which Putnam himself has been a principal advocate, that the world encompasses various *natural kinds*. The apparent tension here runs parallel to that noted earlier, between Putnam's advocacy of an externalist account, in broadly causal terms, of how reference is 'fixed', on the one hand; and on the other, his insistence that no progress can be made on the problem of explaining how reference can be determinate by appeal to causal relations between our words and appropriate bits of the world. So, unless the tension can be argued to be merely apparent, some qualification is needed. We cannot pursue this somewhat delicate issue here, and must content ourselves with one brief cautionary remark. Even supposing that the repudiation of sortal predetermination needs qualification to make space for belief in natural kinds, it would be a mistake, for at least two reasons, to think that this could be exploited to recover a metaphysically realist conception of determinate reference. First, the hypothesis that certain things instantiate a natural kind would, at best, serve to explain the *unity* of a class of things forming the reference or extension of a predicate – as distinct from explaining what *constitutes* reference to that class. (This point may, we suspect, contain the germ of a resolution of the apparent tension – but that is a further issue.) Second, however precisely the envisaged qualification might run, it would be

restricted in scope in a way which would, even prescinding from the previous point, preclude its yielding a fully general solution to the problem with which Putnam confronts the metaphysical realist. Crucially, we could expect no help with explaining how *non*-natural-kind terms can enjoy determinate reference. Essentially the same limitation vitiates David Lewis's proposal (1984, pp. 226–9) that some things, such as rabbits, are more eligible to be the referents of our words than others, such as undetached rabbit parts, and that rival schemes of reference may be ranked as better or worse to the extent that their assignments of referents respect 'nature's joints'. Indeed, the difficulty is not just that appeal to natural divisions could afford an at best partial solution to the general problem; it can be seen, on reflection, that it fails to accomplish even that much – the permutation argument can just as well work to deliver perverse jigglings of perfectly *eligible* referents, and has no need for play with unnatural divisions at all.

42 From 'Why There isn't a Ready-Made World' in Putnam (1983, p. 208).

43 Putnam (1989, p. 220).

44 Indeed, Putnam himself has recently shown signs of a cooling in his opposition to realism as Dummett conceives it (e.g. 1994b, pp. 503 and 510–11).

45 We are indebted to Philip Percival, and to colleagues who attended the Putnam conference in Utrecht in September 1994, especially Putnam himself.

46 The point of this complication is simply to avoid making the needlessly restrictive – and unrealistic – assumption that possible worlds do not differ in point of which objects they contain. In the special case where that assumption holds, we could dispense with the function σ, and need only consider a single permutation ϕ of the domain common to all possible worlds. This special case is, of course, covered by Theorem 2 as stated.

47 Each of the permutations ϕ_i could, of course, be defined to be the restriction to $\sigma(w_i)$ of a single permutation ϕ of the inclusive set D.

References and further reading

Works by Hilary Putnam

1975: The meaning of 'Meaning'. In Putnam's *Mind, Language and Reality: Philosophical Papers*. Vol. 2, Cambridge: Cambridge University Press.

1977: Models and reality. In Putnam (1983).

1978: *Meaning and the Moral Sciences*. Boston and London: Routledge and Kegan Paul.

1981a: *Reason, Truth and History*. Cambridge: Cambridge University Press.

1981b: Beyond historicism. In Putnam (1983).

1981c: Why there isn't a ready-made world. In Putnam (1983).

1983: *Realism and Reason: Philosophical Papers*. Vol. 3, Cambridge: Cambridge University Press.

1988: *Representation and Reality*. Cambridge, Mass.: Bradford/MIT Press.

1989: Model theory and the 'Factuality' of semantics. In Alexander George (ed.), *Reflections on Chomsky*. Oxford: Blackwell.

1990a: A defense of internal realism. In (1990b).

1990b: *Realism with a Human Face*. Cambridge, Mass.: Harvard University Press.

1992: *Renewing Philosophy*. Cambridge, Mass.: Harvard University Press (based on the Gifford Lectures given at St Andrews in 1990).

1994a: Simon Blackburn on internal realism. In Peter Clark and Bob Hale (eds), *Reading Putnam*. Oxford: Blackwell.

1994b: Sense, nonsense and the senses: an enquiry into the powers of the human mind. *Journal of Philosophy*, 91, 445–517 (based on the John Dewey Lectures given at Columbia University in 1994).

Works by other authors

Simon Blackburn 1994: Enchanting views. In Peter Clark and Bob Hale (eds), *Reading Putnam*. Oxford: Blackwell.

Paul Boghossian 1991: Naturalizing content. In B. Loewer and G. Rey (eds), *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

Anthony Brueckner 1984: Putnam's model-theoretic argument against metaphysical realism. *Analysis*, 44(3), 134–40.

Michael Devitt 1983: Realism and the renegade Putnam. *Noûs*, 17, 291–301.

Gareth Evans 1973: The causal theory of names. *Aristotelian Society Suppl. Vol.*, 47, 187–208.

Hartry Field 1972: Tarski's theory of truth. *Journal of Philosophy*, 69, 347–75.

Jerry Fodor 1990: *A Theory of Content*. Cambridge, Mass.: MIT Press.

Ian Hacking 1983: *Representing and Intervening*. Cambridge: Cambridge University Press.

David Lewis 1984: Putnam's paradox. *Australasian Journal of Philosophy*, 62(3), 221–36.