9
Hilary Putnam
Reason, Truth and History

Peter Clark

In the late 1970s and early 1980s Hilary Putnam produced a major sequence of philosophical works all directed at criticism of a certain view of the relation between language and reality. Two of the most salient of those works were *Reason, Truth and History* (1981; hereafter *RTH*) and *Meaning and the Moral Sciences* (1978). Both works were independently philosophical *tours de force* and both were enormously influential, producing a huge secondary literature. This essay concerns principally the former work, although we shall often have to refer to the latter also. Putnam is unselfconsciously one of those philosophers[1] who is not afraid to change his mind and although he now no longer accepts one of the positive claims of *Reason, Truth and History*, namely internal realism (of which much later), the lasting significance of this work is the nexus of philosophical considerations, particularly concerning the notion of reference, which were raised in the book. These considerations are breathtaking in scope, ranging from a refutation of Cartesian scepticism, through numerous insights in the history of philosophy, to issues concerning the theory of truth and the proper interpretation of well-known limitative theorems in mathematical logic. However, the work should not be thought of as a narrow work in analytic philosophy for not only is it replete in allusions to what is called the "continental tradition" in philosophy but Putnam constantly returns to the notion of the "life-enhancing", to the notion of human flourishing and this book systematically exhibits the enormous humanitarian and social concern that motivates so much of his thought.

Putnam announces in his preface to *RTH* that his major concern is to undermine certain traditional dichotomies both of common sense and traditional philosophy, which he argues are unfounded and deeply misleading. Among these are the mind and the world, the objective and the subjective view of truth and reason, and of fact and value. These dichotomies, he argues, are ill defined and misleading but they are all consequences of a deeply held, very influential, but fundamentally mistaken metaphysical view, that of metaphysical realism. His book is a sustained attempt to show the untenability of this view and to replace it with a radically different thesis, which he calls "internal realism". Once internal realism is accepted the untenable dichotomies no longer follow. The cognitive and moral alienation induced by conceiving the world according to metaphysical realism as existing totally independently of our conceptual apparatus, and thus devoid of value, is replaced by a much superior understanding of our place in nature and of the character of knowledge and truth.

Putnam articulates two contrasting philosophical perspectives: that of the externalist and that of the internalist. He characterizes the externalist perspective as follows:

> On this perspective, the world consists of some fixed totality of mind-independent objects. There is exactly one true and complete description of "the way the world is". Truth involves some sort of correspondence relation between words or thought-signs and external things and sets of things. I shall call this perspective the *externalist* perspective, because its favorite point of view is a God's Eye point of view. (*RTH*: 49)

On the other hand, the view he wishes to defend, the internalist perspective, holds that:

> *what objects does the world consists of?* is a question that it only makes sense to ask *within* a theory or description. Many "internalist" philosophers, though not all, hold further that there is more than one "true" theory or description of the world. "Truth", in an internalist view, is some sort of (idealized) rational acceptability – some sort of ideal coherence of our beliefs with each other and with our experiences *as those experiences are themselves represented in our belief system* – and not correspondence with mind-independent or discourse-independent "states of affairs". There is no God's Eye point of view that we can know or usefully imagine; there are only various points of

> view of actual persons reflecting various interests and purposes that their descriptions and theories subserve. (*RTH*: 49–50)

He later reiterates the point concerning theory dependence and the role of conceptual schemes from the internalist perspective:

> In an internalist view also, signs do not intrinsically correspond to objects, independently of how those signs are employed and by whom. But a sign that is actually employed in a particular way by a particular community of users can correspond to particular objects *within the conceptual scheme of those users*. "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. (*RTH*: 52)

Characteristic of the external perspective is the doctrine of metaphysical realism. But what exactly is metaphysical realism? It is not entirely straightforward to say, as one might expect with so pervasive and deep a view. It might be best to approach it metaphorically at first and then to try to do better with a specific philosophical claim. We shall follow Putnam's conception that the view is closely associated with the "God's Eye" perspective.

A little philosophical fantasy

There is a stunning relief etching with watercolour by William Blake completed in 1794 entitled *Ancient of Days*. It shows God about the design and creation of the world. God, in the guise of a naked, human male, holds in his hand a pair of protractors and is bent over, deep in thought, using the protractors to mark out the geometry of the world. In the watercolour the language of creation is Euclidean geometry (illustrated by the protractors) and no doubt the laws of creation are those of Newtonian mechanics and the universal law of gravitation, all given expression in the language of the differential calculus. It is as if in God's mind there is a blueprint for the universe and the language of the blueprint is the differential calculus and Euclidean geometry. The planets are all placed in their elliptical orbits moving against a background of absolute space and time in which all the atoms of the universe have been distributed in accordance with this blueprint. So in effect we can think of the blueprint as a set of four constraints: a space–time framework, Newtonian absolute space and time; a distribution of

matter and energy within that framework; the specification of the four fundamental laws of nature of mechanics and gravitation; and finally the laws that govern the combination of atoms (chemistry and biology).

Now let us think of ourselves as observers and scientists in this Newtonian universe. The first thing to notice is that the language of science, the language that essentially we do science in, is the differential calculus and Euclidean geometry. That is also the language of the blueprint. So when we are thinking about the nature of the world there is a pre-established harmony between the way the world is (as is given in the blueprint) and the language of thought about the world. Now of course merely because we speak or think in the language in which the blueprint is written does not mean that what we say is true, but it does mean that what we say will be true or false, just in case it matches the blueprint or not. The world has a definite determinate structure given by the blueprint, and that structure is directly reflected by the language of the blueprint, geometry and calculus. But the language in which we think, in which we do our science, is geometry and the calculus, so the language of thought and the "language of the world" are identical. One, admittedly metaphorical, way of thinking about the claim of metaphysical realism is that there is a "language of the world" in the above sense (a privileged language in which the blueprint of the universe is written) and it is the same as the language of thought or science.

In a sense we might regard the epistemic condition of observers in such a world as epistemically ideal. Although they may formulate false theories, there is a notion of closeness to the truth for such theories, namely how closely they match the design statements in the blueprint, which are formulated in the same language. (The notion of verisimilitude is notoriously language-dependent.) We can imagine that their science, as more and more evidence comes in, will converge towards the statements in the blueprint. Since those statements in the blueprint constitute the exact truth, there is one true account towards which they are aiming: the "theory of everything" as given by the blueprint. Indeed, we can press this fortunate state of affairs much further. We have been concentrating on general claims about the structure of the universe, but we can be much more specific. We can imagine the language of the blueprint extended in such a way as to contain the names of the natural kinds that occur in the universe (in our Newtonian model world this would be a list of the permitted stable combinations of atoms that might arise chemically and biologically, e.g. gold, radium, mammal, bird etc.). This would be the list of the real natural kinds. Our thinkers would succeed in referring to a natural kind using the term *X* just when the extension of the term *X* coincides with the extension of the corresponding natural-kind term in the language of the blueprint in the actual universe and in all possible worlds. Thus our word "Tree" refers to the natural kind it does precisely because there is a blueprint language

term ("tree") which has exactly the extension it does in the actual and all possible worlds. Again, of course, thinkers in our Newtonian model world might be mistaken in thinking that they had picked out a natural kind. They might well think that they had succeeded in referring to the kind "phlogiston", but the blueprint contains no kind coextensive with the substance of heat. Rather, what it is for us to succeed in referring to, say, "water" is precisely for there to be a natural kind in the blueprint the extension of which is all the H_2O molecules and it is exactly that collection that we refer to when we use the term "water".

On the face of it then it looks as if thinkers in such a world are in a more or less epistemically ideal situation: they inhabit a world made up of a unique domain of objects and kinds, with a unique structure specified again by the blueprint. They speak a language coincident with the language of the blueprint, so everything they say is either true or false as to whether it corresponds or does not correspond to the unique structure given by the blueprint. That is roughly the claim of metaphysical realism. Our world may be very different from the Newtonian fantasy in fact, but not in the matter of how language and thought match reality. There is a language-independent reality; the structure of that language-independent reality is nevertheless reflected exactly by the structure of our language, such that each sentence of that language is true just in case what it says corresponds with that reality. As we quoted above, that is exactly Putnam's way of characterizing this view.[2]

Let us return to the thought that observers in our Newtonian fantasy universe find themselves in an ideal epistemic situation. It certainly looks as if they might because thought and reality naturally match each other. If they had really taken in all the data, collected all the evidence and made no inductive mistakes, would they not know the whole truth about their world? Put another way, would their final science, their theory of everything at the end of the process of data-gathering, not be identical with the blueprint – they would know the whole truth and nothing but the truth? However, for what we might call local and global reasons, this could not be the case. To make this point we can start with rather local reasons. Recall that our model universe is Newtonian and so observers in that universe will find it impossible to distinguish on the basis of any data as to whether the world they inhabit is at rest with respect to absolute space or moving with respect to it at a constant non-zero velocity. This paradigm example of Quinean underdetermination (see Quine 1960) of theory by data is not generated by the accident that the model world is Newtonian. The point is generic; for if we ask ourselves what our observers might come to believe about their world we can see that a disastrous epistemic possibility has opened up for such thinkers – that of universal scepticism. We have already noted that in virtue of Quinean underdetermination, even their ideal theory, formed when all the

data are in, might very well be wrong or seriously incomplete. But the thought must occur to our observers that this possibility once admitted will globalize to include all their theories and representations, and may well infect the adequacy of the concepts they employ.

The sceptical possibility arises that all their thought is mismatched with reality. It may very well appear to them to be internally coherent; further, as far as observable matters are concerned it may well appear true. But how do they know, indeed how could they come to know, that it matches the blueprint? The point is they cannot know, argues Putnam, because of their conception of reference and truth implied by their acceptance of metaphysical realism or the "God's Eye" point of view. For all they know they could be brains in a vat, creatures with a rich cognitive life, that is coherent in itself and satisfied by their world of mental representations, but that corresponds not at all to reality. But, Putnam argues, this possibility that the cognitive life of thinkers might bear no resemblance to reality is self-defeating in much the same way that the thought "I do not exist" is when thought by me. So metaphysical realism entails a proposition (the proposition that: it is a real possibility that our best grasp of the way the world is may bear no relation to the nature of that reality) that is false (because it entails its own negation), so metaphysical realism is false.

The general structure of Putnam's claim has been very well put by Wright (1994). It is worth quoting at length. Wright writes:

> It [metaphysical realism] involves thinking of the world as set over against thought in such a way that it is only by courtesy of a deeply contingent harmony, or felicity, that we succeed, if we do, in forming an overall picture of the world which, at least in its basics, is correct. This is what commits the metaphysical realist to the possibility that even an ideal theory might be false or seriously incomplete. And the same kind of thinking surfaces in the idea that the world comes prejointed, as it were, into real kinds, quite independently of any classificatory activity of ours. Once one thinks of the world in that way, one is presumably committed to the bare possibility of conceptual creatures naturally so constituted as *not* to be prone to form concepts which reflect the real kinds that there are. The real character of the world and its constituents would thus elude both the cognition and the comprehension of such creatures.
>
> Putnam's brains in a vat are exactly such creatures: minds doomed by the character of their interaction with the world they inhabit, and by the nature of that world, not to have the concepts they need in order to be able to capture in thought that world's most fundamental

> features and the nature of their relationship with it … Metaphysical realism is committed to the possibility of a certain kind of dislocation, or uncrossable divide between reality and our cognitive activity. If that possibility were realised, there would accordingly, have to be some correct, specific account of the way in which it was realised. And that is just to say that something like the brain in-the-vat story would have to be true. (Wright 1994: 238)

The brain in the vat story

What is the brain in the vat story and why is it self-refuting? The brain in the vat story is simply an exemplification of the sceptical possibility discussed above: in other words, an account of a possible world in which the sceptical possibility is apparently realized. In this world there are thinkers who have a rich cognitive life, communicate in a language superficially very much like English (call it BIVese) and have pure mental representations much like ours. However, they are in fact disembodied brains in a vat and their thoughts correspond in no way to their real condition. Suppose they try in BIVese to formulate the hypothesis that they are indeed brains in a vat. They will say in BIVese "we are brains in a vat", but the expression of BIVese "brains in a vat" cannot possibly refer to brains in a vat. It cannot do so because, by hypothesis, the very causal relations that must obtain between thinkers using the referring expression "brains in a vat" and actual brains and actual vats do not obtain in the case of the envatted thinkers. So whatever, if anything, "brains in a vat" in BIVese refers to, it is not actual brains and actual vats. So were we to formulate this hypothesis while being brains in a vat, we would not actually be formulating the intended thought at all (we would be formulating what Putnam calls "a thought in a merely bracketed sense" (*RTH*: 28) – a sort of pure mental representation). Hence the claim "We are brains in a vat" formulated in BIVese would be in a certain sense self-refuting, since it cannot under the hypothesis that we are brains in a vat formulate the intended thought. Wright (1994: 224) has provided a short formulation of the argument:

(i) Our language is disquotational (that is meaningful expressions refer in the standard way, "cat" refers to cat, etc.).
(ii) In BIVese "brain in a vat" does not refer to brains in a vat.
(iii) In our language "brain in a vat" is a meaningful expression.
(iv) In our language "brain in a vat" refers to brains in a vat (using (i) and (ii)).
(v) So our language is not BIVese (using (iv) and (ii)).

(vi) If we are brains in a vat then our language, if any, is BIVese.
(vii) So we are not brains in a vat (using (v) and (vi)).

Clearly (i) and (ii) are crucial premises. Premise (ii) hinges on not the acceptance of a causal theory of reference but rather the minimal claim that in order for there to be successful reference there must be at least some appropriate causal connection between tokens of the referring term and the objects referred to, although this indeed may be very indirect. In the case in question the hypothesis itself, that we are brains in a vat, effectively rules out there being causal connections of the appropriate sort, for if we are brains in a vat then there are no vats of the right sort for us to be in causal connection with.

Such, then, is the core of Putnam's ingenious and intriguing argument. If metaphysical realism is true, then a certain possibility seems naturally to arise, but entertaining the hypothesis that that possibility holds shows in fact that there can be no such coherent possibility, so metaphysical realism is false. As Putnam puts it the argument is very simple: "So, if we *are* Brains in a Vat, we cannot *think* that we are, except in the bracketed sense [we are Brains in a Vat]; and this bracketed thought does not have reference conditions that would make it *true*. So it is not possible after all that we are Brains in a Vat" (*RTH*: 50–51). As we noted above the core of the argument lies in premises (i) and (ii) so there must be something fundamentally inconsistent among these premises and metaphysical realism. What that inconsistency is is brought out by the model-theoretic arguments.

The model-theoretic arguments

There are in fact two kinds of model-theoretic arguments deployed by Putnam. One is based on a "permutation" argument and the other, in a way by far the most profound, is an argument using the Löwenheim–Skolem theorem (Skolem [1920] 1967)). Again, it is how the metaphysical realist sees successful reference as being achieved that will be at the core of the issue. All thought or mental representation is object directed: all thought is about something. To put it another way, thoughts have the property of intentionality; they characteristically refer to something else. How does the language in which our thoughts are formulated achieve this? How is it possible, asks Putnam, that we are capable, where we are, of achieving successful reference? "How is intentionality, reference, possible?" (*RTH*: 2), he argues, is the real problem.

The view that it is something about the thinker's pure mental state that fixes the reference of his terms was decisively refuted by a central argument of Putnam's paper "The Meaning of 'Meaning'" (see § Further reading) and his *Meaning and*

the Moral Sciences, the famous "Twin Earth" thesis. A speaker on Earth may use the term water to refer to the liquid H_2O, but on Twin Earth a speaker in the exactly the same mental state may refer to a liquid with all the same observable properties but that is not H_2O by the term "water". Then the term "water" used on Twin Earth refers not to water but to another liquid, yet the mental states of both thinkers are, in all relevant senses, exactly the same.

The suggestion that is Putnam's target in *RTH* is the conception that the reference of terms occurring in sentences can be fixed by the truth of whole sentences containing those terms. The idea is a very natural one. Suppose you are trying to explain to someone, who has never met the notion before, what the term "gene" refers to. You might very well tell him all the key molecular, biological and evolutionary facts that genes are supposed to explain and then say that "gene" refers to exactly those objects that in nature make all of these claims true. Now, whether there are any such objects is a matter for nature to determine. After all, as we have already noted there is no substance phlogiston, but that is because it is in fact impossible to make all of the claims characterizing phlogiston actually true together. It just turns out that the truth-conditions for all the claims characterizing phlogiston are not satisfied in nature. The view under discussion says only that if a term has reference then the reference is fixed by giving the truth-conditions of the sentences containing it. Another way of putting the claim is to go back to the notion that all thought is about something. When we express our thoughts we have an intended interpretation in mind; we mean something; we intend to say something. How can we fix the intended interpretation? According to the view in question we can fix the intended interpretation by laying down the constraint that all that we say is true. Now it might be objected that this is an absurd view because it entirely neglects what Putnam himself was at pains to point out: that there are other constraints on reference. He calls these "theoretical and operational" constraints. An operational constraint would be the requirement that we should get the observational data correct, so all the sentences describing experimental data must come out true. An example of a theoretical constraint might be that we pick the simplest theory that does this. So the operational and theoretical constraints together determine which sentences are true and thus the references of the terms in those sentences. But this objection misses the depth of Putnam's insight. What he noted was that the theoretical and operational constraints amounted in fact just to adding more theory, just more sentences that have to be true on the view in question (Putnam calls it the "received view"). So the objection does not carry weight after all. As he puts it:

> The difficulty with the received view is that it tries to fix the intensions and extensions of individual terms by fixing the truth-conditions

> for whole sentences. The idea, as we just saw, is that operational and theoretical constraints (the ones rational inquirers would accept in some sort of ideal limit of inquiry) determine which sentences in the language are *true*. Even if this is right, however, such constraints cannot determine what our terms *refer* to. For there is nothing in the notion of an operational or theoretical constraint to do this directly. And doing it *indirectly*, by putting down constraints which pick out the set of true sentences, and then hoping that by determining the truth-values of whole sentences we can somehow fix what the terms occurring in those sentences refer to, won't work … In fact, it is possible to interpret the entire language in violently different ways, each of them compatible with the requirement that the truth-value of each sentence in each possible world be the one specified. In short, not only does the received view not work; *no view which only fixes the truth-values of whole sentences can fix reference*, even if it specifies truth-values for sentences *in every possible world*. (*RTH*: 32–3)

Why is this so? It is so because of the permutation argument. Let us revise where we are. We have a language L in which is formulated an ideal scientific theory that satisfies all inductive, operational and theoretical constraints. The claim of the "received view" is that the truth of T fixes the reference of all the names and terms in L. The permutation argument simply says: this cannot be the case because of a (the) basic theorem of model theory that isomorphic interpretations of a language satisfy or make true exactly the same sets of sentences. An interpretation of a language is simply an assignment of objects in a domain to the terms and variables of the language, such that when predicates and relations in the language are interpreted as subsets of the domain, the sentences of the language have a truth-value in that domain. A model of a theory is an interpretation of the language of the theory in which all the sentences of the theory have the truth-value true. The basic theorem says that isomorphic models make the same sentences true. An interpretation A of the language L is isomorphic to an interpretation B if essentially A and B have the same structure and are equinumerous with each other, that is if one is a "mirror image" of the other. This notion can be made quite precise. A permutation of a domain is simply a mapping of the domain onto itself that is non-trivial (i.e. we will exclude the identity mapping). So if, for example, our domain A was the set $\{a, b, c\}$, a permutation of the domain is given by the map $f; A$ onto A by $f(a) = b, f(b) = c$ *and* $f(c) = a$. Now f is here a permutation, so the original domain and the permuted domain have exactly the same number of members. Now we can begin to see the force of the permutation argument. Let us take a simple example to make the point.

Go back to our ideal language L. Let us formulate in L the theory T that says of some predicate R of L the following:

> Not everything has R.
> Something has R.
> If anything is identical with u then it does not have R.
> If anything is identical with v it does not have R.

(where u and v are names in L). Let us lay down that these sentences be true. If we assign to the name u in L the object a in A, and to the name v the object c, and we assign to the predicate R of L the subset $\{b\}$ of A, then indeed all the sentences of T come out true. Not everything has R because in A, a and c do not. Something has R because b does and since u is assigned a and v is assigned c in A the remaining two sentences are true. Have we then uniquely determined that R refers to $\{b\}$? We have not. Look at the permuted domain $f[A]$. Now assign to the name u of L the object $f(a)$, that is, b and to the name v the object $f(c)$, that is, a. Assign to R the subset $\{f(b)\}$, that is, $\{c\}$. Not everything has R because a and b do not, and so on. Under this permuted interpretation all the sentences of T are again true. But now the reference of R is $\{c\}$. The question "What does R refer to in A?" cannot be uniquely answered.[3] So simply laying down the constraint that the sentences of T must be true (in A) will not fix the references of the terms in the sentences. In general there will always be isomorphic models that satisfy the same sets of sentences.[4] That is the force of the permutation argument. Putnam says of it: "It follows that there are always infinitely many different interpretations of the predicates of a language which assign the 'correct' truth-values to the sentences in all possible worlds, *no matter how these 'correct' truth-values are singled out*" (RTH: 35). But it should be noted that the italicized phrase in this quote holds only if the singling out is done by the addition of more and more sentences that have to be true – more theory as we saw above.

There is also a second argument that shows the depth of Putnam's attack on the received view, which emerges again from model-theoretic considerations in the context of set theory. It might be thought that such an argument would have only very local significance, perhaps for the philosophy of mathematics alone, but this is not so. To see that it is not so one merely has to reflect on the centrality and significance of set theory (the theory of arbitrary collections or aggregates of objects) in our conceptual scheme and how much of mathematics and physics is embedded in, or reconstructed in, the framework of set theory. In a certain sense set theory is the ideal theory for doing mathematics. Further, the argument involves the crucial notions of "admissible" or "intended" interpreta-

tion and how such a notion can be made intelligible without the postulation of mysterious cognitive powers possessed by the speakers of a language. Essentially Putnam's argument from the Löwenheim–Skolem theorem encapsulates a dilemma that is quite ubiquitous if one tries to understand how an intended interpretation of a theory can be grasped from a metaphysical realist viewpoint: that dilemma is that there is no stable account that does not either collapse into relativism on the one hand or require the postulation of special very mysterious cognitive powers of intuition on the other (see Putnam 1983).

It is a fundamental result of set theory, perhaps the fundamental result of set theory, that the collection of all subsets of the set of natural numbers, although infinite, cannot be put into one-to-one correspondence with the set of all natural numbers itself.[5] More generally, on a very natural account of size or cardinality[6] the cardinality of a set is strictly less than the cardinality of the set of all the subsets of that set. This is very clear in the finite case. If the set A has two members (say A is the set $\{a, b\}$) then it has four subsets: the empty set $\varnothing$ (which is trivially a subset of every set), $\{a\}$, $\{b\}$, and $\{a, b\}$ (again trivially a set is always a subset of itself). The map that takes member a of A to $\{a\}$ and b of A to $\{b\}$ is a one-to-one correspondence from A into the proper subset $\{\{a\}, \{b\}\}$ of $\{\varnothing, \{a\}, \{b\}, \{a, b\}\}$, but there is no one-to-one correspondence from a four-membered set into a two-membered set. Cantor's beautiful theorem shows how to extend this sort of reasoning to the infinite case. If we say that a set is countable if and only if it can be put into one-to-one correspondence with a subset of the natural numbers then it is a fundamental result of set theory that the power set of the natural numbers (that is the set of all subsets of the natural numbers) is uncountable: there are infinite sets that are uncountable.

Now set *theory* is precisely that: it is a theory expressed as a set of postulates or axioms laying down the existence of certain sets and identity conditions for those sets. In the standard textbook formulation of Zermelo–Fraenkel set theory (the mathematical paradigm formulation of set theory) there are some nine axioms,[7] which assert the existence of certain sets and the identity conditions for sets (e.g. there is an infinite set; given any set, the set of all its subsets exists; any two sets are identical if and only if they have exactly the same members). These axioms can be written down in a first-order language, that is, a language that quantifies only over objects. This is very natural since sets are objects and the axioms taken together characterize our notion of a set. But it is just at this point that a difficulty appears. The axioms of set theory are expressed in a first-order language and it is a central result of the model theory of first-order languages that any set of first-order sentences that has an infinite model has a countably infinite model, that is, if there is an interpretation of the set of sentences that makes all of them true and that is infinite, then there is an interpretation the domain of which forms a

collection of objects that can be put into one-to-one correspondence with the natural numbers. This is the downward Löweheim–Skolem theorem (which itself is provable within set theory together with the axiom of choice). But now we appear to have a paradox, a contradiction sometimes called Skolem's paradox. The standard model of set theory (the way we think of the universe of sets) contains the power set of the set of natural numbers as an object. Any model of the axioms must satisfy the theorems of set theory, since they are logical consequences of the axioms. So Cantor's theorem must be true in that model. So in that model the power set of the natural numbers forms an uncountable collection. But by the downward Löwenheim–Skolem theorem, given that set theory has a model, it must have a countable model. But being a model of the theory it must make Cantor's theorem true, so whatever serves in that countable model to represent the power set of the natural numbers must be a countable collection, since the entire domain is countable. But that looks like saying, depending on which interpretation we pick, that the power set of the natural numbers is either countable or uncountable. Which are they?

That this is not a paradox can easily be seen if we deploy what is sometimes called the "outside/inside" account. Although it is true that from the perspective of the standard interpretation of the universe of sets the model provided by the downward Löwenheim–Skolem theorem is countable, and so the object corresponding to the power set of the natural numbers in that model is again countable, there is no object (no function), no set in the domain of that model that counts the object corresponding to the power set of the natural numbers in that model. So it remains entirely true from "inside" the model, so to say, that the power set of the natural numbers is uncountable and so Cantor's theorem is satisfied. Although looked at from the "outside" (the "true" universe of sets) that is a countable model. All sense of contradiction vanishes when the "inside/outside" perspective is understood. As Putnam puts it "What is a 'countable' set from the point of view of one model may be an uncountable set from the point of view of another model" (Putnam 1983: 2).[8]

However, and it was Putnam's insight to see the depth of the matter, a residual issue remains. For it looks as though we are now committed to an ineliminable, perspectival relativism about the notion of set.[9] Ask the question: which is the right perspective? Are we to think of sets in the way given by the standard interpretation or do we think of the universe of sets as provided by the model given by the downward Löwenheim–Skolem theorem? Well, clearly our notion of set is encapsulated by the axioms of set theory, so it might be thought that we could eliminate any relativism by adding more and more axioms, so continually refining the notion of set and thus eliminating non-standard interpretations. But clearly this will not succeed since we will have more and more first-order

sentences that will still be subject to the downward Löwenheim–Skolem theorem and so have ineliminable non-standard (non-standard because countable) interpretations at every stage. So adding more axioms will not solve the problem, but then, as Putnam remarked, "But if axioms cannot capture the 'intuitive notion of a set', what possibly could?" (*ibid*.: 3). It looks as if to avoid the relativism about sets we would have to postulate some special faculty of mathematical intuition that allowed us to grasp what we really have in mind when we talk about sets in a way that is not linguistically communicable in its entirety. But this seems a hopeless cause.

As Putnam says, the argument from the downward Löwenheim–Skolem theorem can be extended, just as the permutation argument can, to the whole of our corpus of beliefs. It amounts again to the point that adding further sentences expressing further constraints will not fix reference. It is worth quoting him at length on the point:

> Now the argument that Skolem gave, and that shows that "the intuitive notion of a set" (if there is such a thing) is not "captured" by any formal system, shows that even a *formalization of total science* (if one could construct such a thing), or even a *formalization of all our beliefs* (whether they count as "science" or not), could not rule out denumerable interpretations, and, *a fortiori*, such a formalization could not rule out *unintended* interpretations of this notion.
>
> This shows that "theoretical constraints", whether they come from set theory itself or from "total science", cannot fix the interpretation of the notion *set* in the "intended" way. What of "operational constraints"?
>
> Even if we allow that there might be a *denumerable infinity* of measurable magnitudes, and that each of them might be measured to *arbitrary rational accuracy* … it wouldn't help … In short, there certainly seems to be a *countable* model of our *entire body of belief* which meets all operational constraints.
>
> The philosophical problem appears just at this point. If we are told "axiomatic set theory does not capture the intuitive notion of a set", then it is natural to think that *something else* – our "understanding" – does capture it. But what can our "understanding" come to, at least for a naturalistically minded philosopher, which is more than *the way we use our language*? And the Skolem argument can be extended, as we have just seen, to show that the *total use of language* (operational plus theoretical constraints) does not "fix" a unique "intended interpretation" any more than axiomatic set theory by itself does. (*Ibid*.: 3–4)

There are two possible objections to Putnam's reasoning that might at first seem devastating. One is that the downward Löwenheim–Skolem theorem applies only to first-order languages, that is, those that quantify only over objects. It fails for second-order and higher-order languages, for example, those that permit quantification over properties and relations. It may thus seem that all Putnam's argument amounts to is a *non sequitur*; since the axioms of set theory can be given a second-order formulation, why then insist on a first-order formulation? Further, models of set theory in its second-order formulation are unique up to isomorphism so the problem of non-isomorphic interpretations that arises with the downward Löwenheim–Skolem theorem would not appear. But this objection will not work for it simply reintroduces the problem in another way. The problem will re-emerge because we now have to understand how to interpret quantification over arbitrary properties and that really means we will have to be presumed to have a prior grasp of the notion of an arbitrary subset of a set and that in the end will be subject to the same relativism as our first-order notion of a set. Non-isomorphic models of set theory will certainly exist if we do not allow quantification over the full power set of the set of all individuals. Thus, the move to second-order languages will not eliminate the fundamental dilemma. A second and rather more telling objection is that the best that the argument can do is to show that even if we add "total science" to the whole of set theory – that is, add every theoretical and operational constraint we may wish to set theory – we will have no *guarantee* that we will thereby have fixed a unique interpretation for the fundamental notion of set. But this is of no help to the metaphysical realist, for as long as unintended interpretations might be available the general enterprise of metaphysical realism – to show how language succeeds in referring, because our understanding determines a unique reference by eliminating all unintended ones – is undermined. It is of no use to say that language fixes a unique interpretation, when it is always possible that unintended interpretations may very well exist at all stages of enquiry, even at the limit stage when everything by way of additional constraints expressed as more claims in the language is in.

Indeed, there are further ways in which set theory and metaphysical realism make very uneasy bedfellows. The metaphysical realist wants to think of the universe and so the universe of sets as a definite *object* with a structure. But what sort of object? It cannot be a set, for if it were we could form the subset of it corresponding to the set of all sets that are not members of themselves; but there is no such set on pain of Russell's paradox.[10] It could be thought of as a special sort of object called a (proper) class, but we do not have the slightest idea as to why some classes cannot be sets except that we get a contradiction if we suppose them to be. Further, since every set has a power set, so the

universe of sets is indefinitely extensible, there is nothing that can constitute a natural end to the process of obtaining "new" sets. It is indeed very difficult to think of such a domain as an object that we can grasp in any sense independently of how we understand the axioms of set theory. But that is just what we are required to do by the metaphysical realist. He insists that we are talking about that structure (the universe of sets), but there is no way of saying what that structure is other than by laying down certain sentences (the axioms) as true. But we know that will not fix a unique structure because of the existence of unintended interpretations.

Putnam's diagnosis of the problem was that it stemmed from the fundamental thesis of metaphysical realism that language has to be tied to its intended interpretation by the true reference relation, which really determines what we mean and that comes from thinking of the world as a fixed independently existing structure to be conceived of as entirely independent of our conceptual activity. He believes that this commits us to an insoluble dilemma, inescapable perspectival relativism or the possession of mysterious cognitive powers to grasp what is never articulated. But the dilemma is an illusion driven by a false view of the relation between language and reality. We shall let him have the last word:

> The problem, however, lies with the predicament itself. The predicament only *is* a predicament because we did two things: first, we gave an account of understanding the language in terms of programs and procedures for *using* the language (what else?); and then, secondly, we asked what the possible "models" for the language were, thinking of the models as existing "out there" independent of any description. At this point, something really weird had already happened, had we stopped to notice. On any view, the understanding of the language must determine the reference of the terms, or, rather, must determine the reference given the context of use. If the use, even in a fixed context, doesn't determine reference, then use isn't understanding. The language on the perspective we talked ourselves into, has a full programme of use; but it still lacks an interpretation.
>
> This is the fatal step. To adopt a theory of meaning according to which a language whose whole use is specified still lacks something – namely its "interpretation" – is to accept a problem which *can* only have crazy solutions. To speak as if *this* were my problem, "I know how to use my language, but, now, how shall I single out an interpretation?" is to speak nonsense. Either the use already fixes the 'interpretation' or *nothing* can. (Putnam 1983: 23–4)[11]

Notes

1. Bertrand Russell is another example of a philosopher not afraid to change his mind. Indeed, Putnam bares a strong resemblance as a philosopher to Russell in at least two respects. Russell was a consummate practitioner and contributor to mathematical logic, as is Putnam, and Putnam like Russell is passionately concerned with social and moral issues.
2. It is certainly true that traditional realism has held to at least four assumptions: (i) there is a fixed totality of all objects – of things that there are; (ii) there is a fixed totality of properties and relations; (iii) within that second totality there is an unambiguous partition between properties we project on to the world (say evaluative and moral properties) and properties intrinsic to the world; and (iv) there is a fixed relation of "correspondence" between statements and the world that is sufficient to define the notion of a true statement.
3. More generally the procedure is as follows. Look at the domain of A. Call it $D(A)$. Let the one-place (for simplicity) relation or predicate R, part of the vocabulary of T, be interpreted in A by the relation RA holding among a non-empty proper subset of the objects in $D(A)$. Let f be a (one–one) permutation of $D(A)$. Then we can define a new one-place relation Rf on the permuted domain $f[D(A)]$ as follows: $Rf(f(a))$ if and only if $RA(a)$ for each a in $D(A)$. Now we have a new interpretation of the language L; its domain is the same but it assigns different objects to at least one one-place relation of the language L. Recall that the one-place relation RA is interpreted as a *proper* subset of the domain of A. So we can arrange for the permutation f to assign to a an object not having the property RA. So Rf will be different from our original R; different objects will fall under it; the reference of R (a predicate in the language L) will be different in the original model (where it is RA) and the permuted one (where it is Rf). Finally, if $<a_1, a_2, \ldots, a_n, \ldots>$ is any sequence of objects of $D(A)$ that, when assigned to the variables of L, make the sentences of T true, simply assign the sequence of objects $< f(a_1), f(a_2), \ldots, f(a_n), \ldots >$. What we can do for one non-trivial relation R occurring in T we can do for all of them together. We can readily see that the two interpretations are isomorphic, essentially because the permutation is one–one, so they will satisfy exactly the same sets of sentences. The new predicate or relation Rf is just what Putnam denotes as the * property. So in his example RA is cat and Rf is cat*. Similarly, if S were another predicate of L, in Putnam's example SA would be mat and Sf would be mat* (see *RTH*: 34–8).
4. The existence of isomorphic models is very important in understanding what our theoretical knowledge can consist in. It is undoubtedly a very awkward phenomenon for various forms of empiricist accounts of our theoretical knowledge. See particularly William Demopoulos, "On the Rational Reconstruction of our Theoretical Knowledge" *British Journal for the Philosophy of Science* **54**(3) (2003), 371–403.
5. A natural number is any member of the unending sequence 0, 1, 2, 3, 4, 5, 6, …
6. We can say that two sets have the same cardinality (or have the same cardinal number) if and only if there is a one-to-one correspondence among their members; that is, two sets have the same cardinality if and only if they are equinumerous. A set A may be said to have a cardinality strictly less than set B, if there is a one-to-one correspondence from A into a proper subset of B but no one-to-one correspondence exists between B and a subset of A.

7. Strictly speaking this is not correct, for two of the "axioms" are actually axiom *schema*, that is they stand for what is an infinite list of axioms. Thus the Zermelo separation schema – which says that for any set x and any condition F formalizable in the language of set theory there is a subset of x whose members are precisely those members of x that satisfy F – is really an infinite list of axioms each one of that list being an axiom for a specific condition F. A second example is the axiom schema of replacement, which says in effect that if x is a set and F any functional condition then the result of applying F to the members of the set x is also a set. This is really an infinite list of axioms, each axiom corresponding to a specific functional condition F.
8. This is also true of such notions as "is finite" or "is the power set of a given set".
9. This is a conclusion that Skolem himself drew in "Some Remarks on Axiomatised Set Theory", translated and reprinted in *From Frege to Gödel: A Source Book in Mathematical Logic*, J. van Heijenoort (ed.), 290–301 (Cambridge, MA: Harvard University Press, [1922] 1967).
10. Consider the condition formalizable in set theory that holds of a given set x if it is not a member of itself. By the Zermelo separation schema mentioned above, if the universe were a set then the collection of all sets that satisfy the condition would itself be a set. So we would have a set, call it r, the members of which are all and only those sets that are not members of themselves. What about r itself? If r is not a member of r then, by the fact that *all* non-self-membered sets are members of r, r must be a member of r. So r is a member of r. But then since something is a member of r only if it is not self-membered, r cannot be a member of r – which is a contradiction. So the universe cannot be a set; if it were we could apply the Zermelo separation schema for the condition "not being self-membered" and get the contradiction.
11. As Putnam himself says (*RTH*: 6, 66–9) there is a very close connection between these considerations and those of Wittgenstein on rule-following in *Philosophical Investigations*, G. E. M. Anscombe & R. Rhees (eds), G. E. M. Anscombe (trans.) (Oxford: Blackwell, 1953), para. 143–242.

References

Demopoulos, W. 2003. "On the Rational Reconstruction of our Theoretical Knowledge". *British Journal for the Philosophy of Science* **54**(3), 371–403.

Putnam, H. 1975. "The Meaning of 'Meaning'". In *Language, Mind and Knowledge*, K. Gunderson (ed.). Minneapolis, MN: University of Minnesota Press.

Putnam, H. 1978. *Meaning and the Moral Sciences*. London: Routledge & Kegan Paul.

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

Putnam, H. 1983. "Models and Reality". Reprinted in *Realism and Reason: Philosophical Papers, Volume 3*, 1–25. Cambridge: Cambridge University Press

Quine, W. V. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Skolem, T. [1920] 1967. "Logico-combinatorial Investigations in the Satisfiability or Provability of Mathematical Propositions: A Simplified Proof of a Theorem by L. Lowenheim and Generalisations of the Theorem". Translated and reprinted in *From Frege to Gödel: A Source Book in Mathematical Logic*, J. van Heijenoort (ed.), 252–63. Cambridge, MA: Harvard University Press.

Skolem, T. [1922] 1967. "Some Remarks on Axiomatised Set Theory". Translated and reprinted in *From Frege to Gödel: A Source Book in Mathematical Logic*, J. van Heijenoort (ed.), 290–301. Cambridge, MA: Harvard University Press.

Wittgenstein, L. 1953. *Philosophical Investigations*, G. E. M. Anscombe & R. Rhees (eds), G. E. M. Anscombe (trans.). Oxford: Blackwell.

Wright, C. 1994. "On Putnam's Proof that we are not Brains-in-a-Vat". In *Reading Putnam*, P. Clark & B. Hale (eds), 216–41. Oxford: Blackwell.

Further reading

The evolution of Putnam's views on realism makes a fascinating study. Some of his early papers were highly critical of challenges to realism. Particularly notable in this respect are his papers "The Refutation of Conventionalism" and "The Meaning of 'Meaning'", reprinted in Hilary Putnam, *Mind, Language and Reality: Philosophical Papers, Volume 2* (Cambridge: Cambridge University Press, 1975), 153–91, 215–71, respectively, and "What is Mathematical Truth?", reprinted in Hilary Putnam, *Mathematics, Matter and Method: Philosophical Papers, Volume 1* (Cambridge: Cambridge University Press, 1975), 60–78. By the early 1980s, however, he had abandoned metaphysical realism and adopted "internal realism". Two classic papers laying out his arguments are "Models and Reality", reprinted in *Realism and Reason: Philosophical Papers, Volume 3* (Cambridge: Cambridge University Press, 1983), 1–25, and "Why There isn't a Ready Made World", reprinted in *Realism and Reason,* 205–28.

Three of the best critical responses to Putnam's "Models and Reality" are those of David Lewis, "Putnam's Paradox", *Australasian Journal of Philosophy* **62** (1984), 221–36, Timothy Bays, "On Putnam and His Models", *Journal of Philosophy* **98**(7) (2001), 331–50, and M. Devitt, *Realism and Truth*, 2nd edn (Oxford: Oxford University Press, 1991), esp. 220–34, 330–38. *Reading Putnam* (Oxford: Blackwell, 1994), the collection of papers edited by Peter Clark and Bob Hale, contains nine papers largely devoted to Putnam's defence of internal realism. The contribution by Michael Hallett, "Putnam and the Skolem Paradox" (*Reading Putnam*, 66–97) is a sustained treatment of the model-theoretic arguments. The background work in logic and set theory to the model-theoretic arguments can be found in Robert R. Stoll, *Set Theory and Logic* (San Francisco, CA: W. H. Freeman, 1961), particularly chapters 5, 7 and 9. Crispin Wright's paper "On Putnam's Proof that we are not Brains-in-a-Vat" (*Reading Putnam*, 216–41) is one of the best treatments in the literature of Putnam's ingenious argument. Two further collections on Putnam's work on realism in general are C. S. Hill (ed.), *The Philosophy of Hilary Putnam*, *Philosophical Topics* **20**(1) and James Conant & Urszula M. Zeglen (eds), *Hilary Putnam: Pragmatism and Realism* (London: Routledge, 2002).

The later development of Putnam's views on realism subsequent to the publication of *Reason, Truth and History* can be found in his Paul Carus Lectures, *The Many Faces of Realism* (La Salle, IL: Open Court, 1987), his Gifford Lectures delivered at the University of St Andrews, *Renewing Philosophy* (Cambridge, MA: Harvard University Press, 1992) and "The Dewey Lectures 1994: Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind", *Journal of Philosophy* **91**(9) (September 1994); all three are collected in Putnam's *The Threefold Cord: Mind, Body and World*, (New York: Columbia University Press, 1999). For his current views on some of the themes in *Reason, Truth and History* that

have not been treated in this essay see his *The Collapse of the Fact /Value Dichotomy and Other Essays* (Cambridge, MA: Harvard University Press, 2004) and his *Ethics without Ontology* (Cambridge, MA: Harvard University Press, 2004). The twin earth argument and some of the central themes of *Meaning and the Moral Sciences* are explored extensively in Andrew Pressin & Sanford Goldberg (eds), *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's the "Meaning of Meaning"* (Armonk, NY: M. E. Sharpe, 1996).