3

Two philosophical perspectives

The problems we have been discussing naturally give rise to two philosophical points of view (or two philosophical temperaments, as I called them in the Introduction). It is with these points of view, and with their consequences for just about every issue in philosophy that I shall be concerned: the question of 'Brains in a Vat' would not be of interest, except as a sort of logical paradox, if it were not for the sharp way in which it brings out the difference between these philosophical perspectives.

One of these perspectives is the perspective of metaphysical realism. 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.

The perspective I shall defend has no unambiguous name. It is a late arrival in the history of philosophy, and even today it keeps being confused with other points of view of a quite different sort. I shall refer to it as the *internalist* perspective, because it is characteristic of this view to hold that *what objects does the world consist 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 the various points of view of actual persons reflecting various interests and purposes that their descriptions and theories subserve. ('Coherence theory of truth'; 'Non-realism'; 'Verificationism'; 'Pluralism'; 'Pragmatism'; are all terms that have been applied to the internalist perspective; but every one of these terms has connotations that are unacceptable because of their other historic applications.)

Internalist philosophers dismiss the 'Brain in a Vat' hypothesis. For us, the 'Brain in a Vat World' is only a *story*, a mere linguistic construction, and not a possible world at all. The idea that this story might be true in some universe, some Parallel Reality, assumes a God's Eye point of view from the start, as is easily seen. For *from whose point of view is the story being told?* Evidently *not* from the point of view of any of the sentient creatures *in* the world. Nor from the point of view of any observer in another world who interacts with this world; for a 'world' by definition includes everything that interacts in any way with the things it contains. If *you*, for example, were the one observer who was *not* a Brain in a Vat, spying on the Brains in a Vat, then the world would not be one in which *all* sentient beings were Brains in a Vat. So the supposition that there could be a world in which *all* sentient beings are Brains in a Vat presupposes from the outset a God's Eye view of truth, or, more accurately, a No Eye view of truth – truth as independent of observers altogether.

For the externalist philosopher, on the other hand, the hypothesis that we are all Brains in a Vat cannot be dismissed so simply. For the truth of a theory does not consist in its fitting the world as the world presents itself to some observer or observers (truth is not 'relational' in this sense), but in its corresponding to the world as it is in itself. And the problem that I posed for the externalist philosopher is that the very relation of correspondence on which truth and reference depend (on his view) cannot logically be available to him if he *is* a Brain in a Vat. 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.

Suppose we assume a 'magical theory of reference'. For example, we might assume that some occult rays – call them 'noetic rays'[1] – connect words and thought-signs to their referents. Then there is no problem. The Brain in a Vat can think the *words*, 'I am a brain in a vat', and when he does the word 'vat' corresponds (with the aid of the noetic rays) to real external vats and the word 'in' corresponds (with the aid of the noetic rays) to the relation of real spatial containment. But such a view is obviously untenable. No present day philosopher would espouse such a view. It is because the modern realist wishes to have a correspondence theory of truth *without* believing in 'noetic rays' (or, believing in Self-Identifying Objects[2] – objects that intrinsically correspond to one word or thought-sign rather than another) that the Brain in a Vat case is a puzzler for him.

As we have seen, the problem is this: there are these objects out there. Here is the mind/brain, carrying on its thinking/computing. How do the thinker's symbols (or those of his mind/brain) get into a unique correspondence with objects and sets of objects out there?

The reply popular among externalists today is that while indeed no sign *necessarily* corresponds to one set of things rather than another, *contextual* connections between signs and external things (in particular, causal connections) will enable one to explicate the nature of reference. But this doesn't work. For example, the dominant cause of my beliefs about electrons is probably various *textbooks*. But the occurrences of the word 'electron' I produce, though having in this sense a strong connection to textbooks, do not *refer* to textbooks. The objects which are the dominant cause of my beliefs containing a certain sign may not be the referents of that sign.

The externalist will now reply that the word 'electron' is not connected to textbooks by a causal chain *of the appropriate type*. (But how can we have intentions which determine which causal chains are 'of the appropriate type' unless we are *already* able to *refer?*)

[1] 'Noetic rays' was suggested to me by Zemach.

[2] The term 'Self Identifying Object' is from *Substance and Sameness* by David Wiggins (Blackwell, 1980).

For an internalist like myself, the situation is quite different. 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.

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! What does 'extraterrestrial' refer to? To extraterrestrials (if there are any).

Of course the externalist agrees that the extension of 'rabbit' is the set of rabbits and the extension of 'extraterrestrial' is the set of extraterrestrials. 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. (*How* pressing, we saw in the previous chapter.) For me there is little to say about what reference is within a conceptual system other than these tautologies. The idea that causal connection is necessary is refuted by the fact that 'extraterrestrial' certainly refers to extraterrestrials whether we have ever causally interacted with any extraterrestrials or not!

The externalist philosopher would reply, however, that we can refer to extraterrestrials even though we have never interacted with any (as far as we know) because we have interacted with *terrestrials* and we have experienced instances of the relation 'not from the same planet as' and instances of the property 'intelligent being'. And we can *define* an extraterrestrial as an intelligent being that is not from the same planet as terrestrials. Also, 'not from the same planet as' can be analyzed in terms of 'not from the same place as' and 'planet' (which can be further analyzed). Thus the externalist gives up the requirement that we have some 'real' connection (e.g. causal connection) with *everything* we are able to refer to, and requires only that the *basic* terms refer to kinds of things (and relations) that we have some

real connection to. Using the basic terms in complex conbinations we can then, he says, build up descriptive expressions which refer to kinds of things we have no real connection to, and that may not even exist (e.g. extraterrestrials).

In fact, already with a simple word like 'horse' or 'rabbit' he might have observed that the extension includes many things we have *not* causally interacted with (e.g. *future* horses and rabbits, or horses and rabbits that never interacted with any human being). When we use the word 'horse' we refer not only to the horses we have a real connection to, but also to all other things *of the same kind*.

At this point, however, we must observe that 'of the same kind' makes no sense apart from a categorial system which says what properties do and what properties do not count as similarities. In *some* ways, after all, anything is 'of the same kind' as anything else. This whole complicated story about how we refer to some things by virtue of the fact that they are connected with us by 'causal chains of the appropriate kind', and to yet other things by virtue of the fact that they are 'of the same kind' as things connected with us by causal chains of the appropriate kind, and to still other things 'by description', is not so much false as otiose. What makes horses with which I have not interacted 'of the same kind' as horses with which I *have* interacted is that fact that the former as well as the latter are *horses*. The metaphysical realist formulation of the problem once again makes it seem as if there are to begin with all these objects in themselves, and then I get some kind of a lassoo over a few of these objects (the horses with which I have a 'real' connection, via a 'causal chain of the appropriate kind'), and then I have the problem of getting my word ('horse') to cover not only the ones I have 'lassooed' but also the ones I can't lassoo, because they are too far away in space and time, or whatever. And the 'solution' to this pseudo-problem, as I consider it to be – the metaphysical realist 'solution' – is to say that the word *automatically* covers not just the objects I lassooed, but also the objects which are *of the same kind* – of the same kind *in themselves*. But then the world is, after all, being claimed to contain Self-Identifying Objects, for this is just what it means to say that the *world*, and not thinkers, sorts things into kinds.

In a sense, I would say, the world *does* consist of 'Self-Identi-

fying Objects' – but not a sense available to an externalist. If, as I maintain, 'objects' themselves are as much made as discovered, as much products of our conceptual invention as of the 'objective' factor in experience, the factor independent of our will, then of course objects intrinsically belong under certain labels; because those labels are the tools we used to construct a version of the world with such objects in the first place. But *this* kind of 'Self-Identifying Object' is not mind-independent; and the externalist wants to think of the world as consisting of objects that are *at one and the same time* mind-independent and Self-Identifying. This is what one cannot do.

Internalism and relativism

Internalism is not a facile relativism that says, 'Anything goes'. Denying that it makes sense to ask whether our concepts 'match' something totally uncontaminated by conceptualization is one thing; but to hold that every conceptual system is therefore just as good as every other would be something else. If anyone really believed that, and if they were foolish enough to pick a conceptual system that told them they could fly and to act upon it by jumping out of a window, they would, if they were lucky enough to survive, see the weakness of the latter view at once. Internalism does not deny that there are experiential *inputs* to knowledge; knowledge is not a story with no constraints except *internal* coherence; but it does deny that there are any inputs *which are not themselves to some extent shaped by our concepts*, by the vocabulary we use to report and describe them, or any inputs *which admit of only one description, independent of all conceptual choices*. Even our description of our own sensations, so dear as a starting point for knowledge to generations of epistemologists, is heavily affected (as are the sensations themselves, for that matter) by a host of conceptual choices. The very inputs upon which our knowledge is based are conceptually contaminated; but contaminated inputs are better than none. If contaminated inputs are all we have, still all we have has proved to be quite a bit.

What makes a statement, or a whole system of statements – a theory or conceptual scheme – rationally acceptable is, in large

part, its coherence and fit; coherence of 'theoretical' or less experiential beliefs with one another and with more experiential beliefs, and also coherence of experiential beliefs with theoretical beliefs. Our conceptions of coherence and acceptability are, on the view I shall develop, deeply interwoven with our psychology. They depend upon our biology and our culture; they are by no means 'value free'. But they *are* our conceptions, and they are conceptions of something real. They define a kind of objectivity, *objectivity for us*, even if it is not the metaphysical objectivity of the God's Eye view. Objectivity and rationality humanly speaking are what we have; they are better than nothing.

To reject the idea that there is a coherent 'external' perspective, a theory which is simply true 'in itself', apart from all possible observers, is not to *identify* truth with rational acceptability. Truth cannot simply *be* rational acceptability for one fundamental reason; truth is supposed to be a property of a statement that cannot be lost, whereas justification can be lost. The statement 'The earth is flat' was, very likely, rationally acceptable 3,000 years ago; but it is not rationally acceptable today. Yet it would be wrong to say that 'the earth is flat' was *true* 3,000 years ago; for that would mean that the earth has changed its shape. In fact, rational acceptability is both tensed and relative to a person. In addition, rational acceptability is a matter of degree; truth is sometimes spoken of as a matter of degree (e.g., we sometimes say, *'the earth is a sphere' is approximately true*); but the 'degree' here is the *accuracy* of the statement, and not its degree of acceptability or justification.

What this shows, in my opinion, is not that the externalist view is right after all, but that truth is an *idealization* of rational acceptability. We speak as if there were such things as epistemically ideal conditions, and we call a statement 'true' if it would be justified under such conditions. 'Epistemically ideal conditions', of course, are like 'frictionless planes': we cannot really attain epistemically ideal conditions, or even be absolutely certain that we have come sufficiently close to them. But frictionless planes cannot really be attained either, and yet talk of frictionless planes has 'cash value' because we can approximate them to a very high degree of approximation.

Perhaps it will seem that explaining truth in terms of justifi-

cation under ideal conditions is explaining a clear notion in terms of a vague one. But 'true' is *not* so clear when we move away from such stock examples as 'Snow is white.' And in any case, I am not trying to give a formal *definition* of truth, but an informal elucidation of the notion.

The simile of frictionless planes aside, the two key ideas of the idealization theory of truth are (1) that truth is independent of justification here and now, but not independent of *all* justification. To claim a statement is true is to claim it could be justified. (2) truth is expected to be stable or 'convergent'; if both a statement and its negation could be 'justified', even if conditions were as ideal as one could hope to make them, there is no sense in thinking of the statement as *having* a truth-value.

The 'similitude' theory

The theory that truth is correspondence is certainly the natural one. Before Kant it is perhaps impossible to find *any* philosopher who did *not* have a correspondence theory of truth.

Michael Dummett has recently[3] drawn a distinction between *non-realist* (i.e. what I am calling 'internalist') views and *reductionist* views in order to point out that reductionists can be metaphysical realists, i.e. subscribers to the correspondence theory of truth. Reductionism, with respect to a class of assertions (e.g. assertions about mental events) is the view that assertions in that class are 'made true' by facts which are outside of that class. For example, facts about behavior are what 'make true' assertions about mental events, according to one kind of reductionism. For another example, the view of Bishop Berkeley that all there 'really is' is minds and their sensations is *reductionist*, for it holds that sentences about tables and chairs and other ordinary 'material objects' are actually made true by facts about sensations.

If a view is reductionist with respect to assertions of one kind, but only to insist on the correspondence theory of truth for sen-

[3] Dummett's views are set out in 'What is a theory of Meaning I, II' in *Truth and Other Enigmas* (Harvard, 1980). His forthcoming (eventually) William James Lectures (given at Harvard in 1976) develop them in much more detail.

tences of the *reducing* class, then that view is metaphysical realist at base. A truly non-realist view is non-realist all the way down.

The error is often made of regarding reductionist philosophers as non-realists, but Dummett is surely right; *their* disagreement with other philosophers is over *what there really is*, and not over the conception of truth. If we avoid this error, then the claim I just made, that it is impossible to find a philosopher before Kant who was *not* a metaphysical realist, at least about what he took to be *basic* or unreducible assertions, will seem much more plausible.

The oldest form of the correspondence theory of truth, and one which endured for approximately 2,000 years, is one that ancient and medieval philosophers attributed to Aristotle. That Aristotle actually held it I am not sure; but it is suggested by his language. I shall call it *the similitude theory of reference;* for it holds that the relation between the representations in our minds and the external objects that they refer to is literally a *similarity*.

The theory, like modern theories, employed the idea of a mental representation. This presentation, the mind's image of the external thing, was called a *phantasm* by Aristotle. The relation between the phantasm and the external object by virtue of which the phantasm represents the external object to the mind is (according to Aristotle) that the phantasm *shares a form* with the external object. Since the phantasm and the external object are similar (share the form), the mind, in having available the phantasm, also has directly available the very *form* of the external object.[4]

Aristotle himself says that the phantasm does *not* share with the object such properties as *redness* (i.e. the redness in our minds is not literally the same property as the redness of the object), which can be perceived by one sense, but does share such properties as *length* or *shape* which can be perceived by more than one sense (which are 'common sensibles' as opposed to 'single sensibles').

In the seventeenth century the similitude theory began to be restricted, much as it had been by Aristotle. Thus Locke and Descartes held that in the case of a 'secondary' quality, such as a color or a texture, it would be absurd to suppose that the prop-

[4] See *De Anima*, Book III, Ch. 7 and 8.

erty of the mental image is *literally* the same property as the property of the physical thing. Locke was a Corpuscularian, that is, an advocate of the atomic theory of matter, and like a modern physicist he conceived that what answers to the sensuous presented redness of my image of a red piece of cloth is not a simple property of the cloth, but a very complex dispositional property or 'Power', namely the Power to give rise to sensations of this particular kind (sensations which exhibit 'subjective red', in the language of psychophysics). This power in turn has an explanation, which we did not know in Locke's day, in the particular micro-structure of the piece of cloth which leads it to selectively absorb and reflect light of different wave-lengths. (This *sort* of explanation was already given by Newton.) If we say that having such a microstructure is 'being red' in the case of a piece of cloth, then clearly whatever the nature of subjective red may be, the event in my mind (or even my brain) that takes place when I have a sensation of subjective red does *not* involve anything in my mind (or brain) 'being red'. The properties of a physical thing which make it an instance of physical red and the properties of a mental event which make it an instance of subjective red are quite different. A red piece of cloth and a red after-image are *not* literally similar. They do not share a Form.

For those properties (shape, motion, position) which his Corpuscularian philosophy led him to regard as basic and irreducible, Locke was willing to keep the similitude theory of reference, however. (Actually, some Locke scholars today dispute this; but Locke does say that there is a 'similitude' between the idea and the object in the case of the primary qualities and that there is 'no similitude' between the idea of *red* or *warmth* and the redness or warmth in the object.[5] And the reading of Locke I am describing was the universal one among his contemporaries and among eighteenth century readers as well.)

Berkeley's tour de force

Berkeley discovered a very unwelcome consequence of the similitude theory of reference: it implies that nothing exists except mental entities ('spirits and their ideas', i.e. minds and

[5] See *An Essay Concerning Human Understanding,* Book II, Ch. VIII.

their sensations). It is generally unappreciated that the premiss from which Berkeley worked – the similitude theory – was not something he merely learned from Locke (or read into Locke) but was the accepted theory of reference before his time and, indeed, for a hundred years afterwards; but we have just remarked how venerable this theory actually was.

Berkeley's argument is very simple. The usual philosophical argument against the similitude theory in the case of secondary qualities is correct (the argument from the relativity of perception), but it goes just as well in the case of primary qualities. The length, shape, motion of an object are all perceived differently by different perceivers and by the same perceiver on different occasions. To ask whether a *table* is the same length as *my* image of it or the same length as *your* image of it is to ask an absurd question. If the table is three feet long, and I have a good clear view of it, do I have a *three foot long mental image?* To ask the question is to see its senselessness. Mental images do not have a *physical* length. They cannot be compared with the standard measuring rod in Paris. Physical length and subjective length must be as different as physical redness and subjective redness.

To state Berkeley's conclusion another way, *Nothing can be similar to a sensation or image except another sensation or image.* Given this, and given the (still unquestioned) assumption that the mechanism of reference is similitude between our 'ideas' (i.e. our images or 'phantasms') and what they represent, it at once follows that no 'idea' (mental image) can represent or refer to anything but another image or sensation. Only phenomenal objects can be thought about, conceived, referred to. And if you can't think of something, you can't think it exists. Unless we treat talk of material objects as highly derived talk about regularities in our sensations, it is totally unintelligible.

The tendency, in his own time and later, to see Berkeley as almost insanely perverse, almost scandalous, if brilliant, was due to the unacceptability of his conclusion that matter does not really exist (except as a construction from sensations), and not to anything peculiar about his premisses. But the fact that one could derive such an unacceptable conclusion from the similitude theory produced a crisis in philosophy. Philosophers who did not wish to follow Berkeley in Subjective Idealism had to come up with a different account of reference.

Kant's account of knowledge and truth

I want to say that, although Kant never quite says that this is what he is doing, Kant is best read as proposing for the first time what I have called the 'internalist' or 'internal realist' view of truth.

To begin with, it is clear that Kant regarded Berkeley's Subjective Idealism as quite unacceptable (this much he explicitly says), and also regarded causal realism – the view that we directly perceive only sensations, and *infer* material objects via some kind of problematical inference, as equally unacceptable. A view on which it is only a very dubious hypothesis that there is a table in front of me as I write these pages is a 'scandal', Kant says.

Secondly, I take it that Kant saw clearly how Berkeley's argument works: he saw that it depends on the similitude theory of reference, and that rejecting Berkeley's argument requires rejecting that theory. Here I am attributing a view to Kant that Kant does not express in these words (indeed, talk of 'reference' as the relation between mental signs and what they stand for is very recent, although the problem of the relation between mental signs and what they stand for is very ancient). But we shall see that what Kant *did* say has precisely the effect of giving up the similitude theory of reference.

Let me suggest a way of reading Kant that may be helpful, although it is only a first approximation to a right interpretation. Think of Kant as accepting Berkeley's point that the argument from the relativity of perception applies as much to the so-called 'primary' qualities as to the secondary ones, but making a different response than Berkeley made. Berkeley's response, recall, was to scrap the distinction between primary qualities and secondary qualities and fall back on just what Locke would have called 'simple' qualities of sensation as the basic entities we can refer to. Locke's own treatment of secondary qualities, recall, was to say that (as properties of the physical object) we can only conceive of them as Powers, as properties – *nature unspecified* – which enable the object to affect *us* in a certain way. Saying that something is red, or warm, or furry, is saying that it is so-and-so *in relation to us*, not how it is from a God's Eye point of view.

I suggest that (as a first approximation) the way to read Kant is as saying that what Locke said about secondary qualities is

true of *all* qualities – the simple ones, the primary ones, the secondary ones alike (indeed, there is little point of distinguishing them).[6]

If *all properties are secondary*, what follows? It follows that *everything* we say about an object is of the form: it is such as to affect *us* in such-and-such a way. *Nothing at all* we say about any object describes the object as it is 'in itself', independently of its effect on *us*, on beings with our rational natures and our biological constitutions. It also follows that we cannot assume any similarity ('similitude', in Locke's English) between our idea of an object and whatever mind-independent reality may be ultimately responsible for our experience of that object. Our ideas of objects are not *copies* of mind-independent things.

This is very much the way Kant describes the situation. He does not doubt that there is *some* mind-independent reality; for him this is virtually a postulate of reason. He refers to the elements of this mind-independent reality in various terms: thing-in-itself (*Ding an sich*); the noumenal objects or *noumena;* collectively, *the noumenal world.* But we can form no real conception of these noumenal things; even the notion of a noumenal world is a kind of limit of thought (*Grenz-Begriff*) rather than a clear concept. Today the notion of a noumenal world is perceived to be an unnecessary metaphysical element in Kant's thought. (But perhaps Kant is right: perhaps we can't help think-

[6] Kant gives a summary of his own view in precisely this way in the *Prolegomena:*

> Long before Locke's time, but assuredly since him, it has been generally assumed and granted without detriment to the actual existence of external things that many of their predicates may be said to belong, not to the things in themselves, but to their appearances, and to have no proper existence outside our representation. Heat, color, and taste, for instance, are of this kind. Now, if I go farther and, for weighty reasons, rank as mere appearances the remaining qualities of bodies also, which are called primary – such as extension, place, and, in general, space, with all that which belongs to it (impenetrability or materiality, shape, etc.) – no one in the least can adduce the reason of its being inadmissible. As little as the man who admits colors not to be properties of the object in itself, but only as modifications of the sense of sight, should on that account be called an idealist, so little can my thesis be named idealistic merely because I find that more, nay, *all the properties which constitute the intuition of a body belong merely to its appearance.*

ing that there is *somehow* a mind-independent 'ground' for our experience even if attempts to talk about it lead at once to nonsense.)

At the same time, talk of ordinary 'empirical' objects is *not* talk of things-in-themselves but only talk of things-for-us.

The really subtle point is that Kant regards all of these points as applying to sensations ('objects of internal sense') *as well as* to external objects. This may seem strange: what is the problem about whether or not an idea corresponds to a sensation? But Kant is on to something profound.

Suppose I have a sensation *E*. Suppose I *describe E;* say, by asserting '*E* is a sensation of *red*.' If 'red' just means *like this*, then the whole assertion just means '*E* is *like this*' (attending to *E*), i.e. *E is like E* – and no judgment has really been made. As Wittgenstein puts it, one is reduced to virtually a grunt. On the other hand, if 'red' is a true *classifier*, if I am claiming that this sensation *E belongs in the same class* as sensations I call 'red' *at other times*, then my judgment goes beyond what is immediately given, beyond the 'bare thatness', and involves an implicit reference to other sensations, which I am not having at the present instant, and to *time* (which, according to Kant, is not something noumenal but rather a form in which we arrange the 'things-for-us').[7] Whether the sensations I have at different times that I classify as *sensations of red* are all 'really' (noumenally) similar is a question that makes no sense; if they appear to be similar (e.g. if I *remember* the previous sensations as similar to this one, and *anticipate* that future sensations which I will so classify will in their turn seem to be similar to this one, as this one is then remembered) then they are similar-for-me.

Kant says again and again, and in different words, that the objects of inner sense are *not* transcendentally real (noumenal) that they are 'transcendentally ideal' (things-for-us), and that they are no more and no less directly knowable than so-called

[7] Here I am being deliberately anachronistic and describing Kant's view by means of an example taken from Wittgenstein's *Philosophical Investigations*. But Wittgenstein's example has deeply Kantian roots: Hegel, writing shortly after Kant, and aware of Kant's doctrine, made precisely the point that any judgment, even of sense impression, has to go beyond what is 'given' to *be* a judgment at all.

'external' objects. The sensations I call 'red' can no more be directly compared with noumenal objects to see if they have the same *noumenal* property than the objects I call 'pieces of gold' can be directly compared with noumenal objects to see if they have the same noumenal property.

The reason that 'All properties are secondary' is only a *first* approximation to Kant's view is this: 'All properties are secondary' (i.e. *all properties are Powers*) suggests that saying of a chair that it is made of pine, or whatever, is attributing a Power (the disposition to appear to be made of pine to us) to a noumenal object; saying of the chair that it is brown is attributing a different Power to that *same* noumenal object; and so on. On such a view there would be one noumenal object corresponding to each object in what Kant calls 'the representation', i.e. one noumenal object corresponding to each thing-for-us. But Kant explicitly *denies* this. This is the point at which he all but says that he is giving up the correspondence theory of truth.

Kant does not, indeed, *say* he is giving up the correspondence theory of truth. On the contrary, he says that truth is the 'correspondence of a judgment to its object'. But this is what Kant called a 'nominal definition of truth'. On my view, identifying this with what the metaphysical realist means by 'the correspondence theory of truth' would be a grave error. To say whether Kant held what a metaphysical realist means by 'the correspondence theory of truth' we have to see whether he had a realist conception of what he called 'the object' of an empirical judgment.

On Kant's view, any judgment about external or internal objects (physical things or mental entities) says that the noumenal world as a whole is such that this is the description that a rational being (one with our rational nature) given the information available to a being with our sense organs (a being with our sensible nature) would construct. In *that* sense, the judgment ascribes a Power. But the Power is ascribed *to the whole noumenal world*; you must *not* think that because there are chairs and horses and sensations in our representation, that there are correspondingly noumenal chairs and noumenal horses and noumenal sensations. *There is not even a one-to-one correspondence between things-for-us and things in themselves.* Kant not only gives up any notion of similitude between our ideas and the

things in themselves; he even gives up any notion of an abstract isomorphism. And this means that there is no correspondence theory of truth in his philosophy.

What then is a true judgment? Kant does believe that we have *objective* knowledge: we know laws of mathematics, laws of geometry, laws of physics, and many statements about individual objects – empirical objects, things for us. The use of the term 'knowledge' and the use of the term 'objective' amount to the assertion that *there still is a notion of truth*. But what *is* truth if it is not correspondence to the way things are in themselves?

As I have said, the only answer that one can extract from Kant's writing is this: a piece of knowledge (i.e. a 'true statement') is a statement that a rational being would accept on sufficient experience of the kind that it is actually possible for beings with our nature to have. 'Truth' in any other sense is inaccessible to us and inconceivable by us. *Truth is ultimate goodness of fit.*

The empiricist alternative

So far as our argument has gone, it is still possible for a philosopher to avoid giving up the correspondence theory of truth and the similitude theory of reference by *restricting them to sensations and images*. And many philosophers continued to believe even after Kant that similitude *is* the mechanism by which we are able to have ideas that refer to our own (and, although this was more controversial, other people's) *sensations*, and that this is the primary case of reference from an epistemological point of view.

To see why this doesn't work, recall that the heart of Berkeley's argument was the contention that nothing can resemble an 'idea' (sensation or image) except another 'idea', i.e. there can be no resemblance between the mental and the physical. Our ideas can resemble other mental entities, but they cannot resemble 'matter', according to Berkeley.

At this point, we must stop and realize that this is in an important way false. In fact, *everything is similar to everything else in infinitely many respects*. For example, my sensation of a typewriter at this instant and the quarter in my pocket are both similar in the respect that some of their properties (the sensation's

occurring right now and the quarter's being in my pocket right now) are *effects of my past actions;* if I had not sat down to type, I would not be having the sensation; and the quarter would not be in my pocket if I had not put it there. Both the sensation and the quarter exist in the twentieth century. Both the sensation and the quarter have been described in English. And so on and so on. The number of similarities one can find between *any* two objects is limited only by ingenuity and time.

In a particular context, 'similarity' may have a more restricted meaning, of course. But to just ask 'are *A* and *B* similar?' when we have not specified, explicitly or implicity, what *kind* of similarity is at issue, is to ask an empty question.

From this simple fact it already follows that the idea that similitude is the private mechanism of reference must lead to an infinite regress. Suppose, to use an example due to Wittgenstein, someone is trying to invent a 'private language', a language which refers to his own sensations as they are directly given to him. He focusses his attention on a sensation *X* and introduces a sign *E* which he intends to apply to exactly those entities which are qualitatively identical with *X*. In effect, he intends that *E* should apply to all and only those entities which are *similar* to *X*.

If this is *all* he intends – if he does not specify the *respect* in which something has to be similar to *X* to fall under the classification *E* – then his intention is empty, as we just saw. For *everything* is similar to *X* in *some* respect.

If, on the other hand, he *specifies* the respect; if he thinks the thought that *a sensation is E if and only if it is similar to X in respect R;* then, since he is able to think this thought, he is *already* able to refer to the sensations for which he is trying to introduce a term *E,* and to the relevant property of those sensations! But how did he get to be able to do *this?* (If we answer, 'By focussing his attention on two other sensations, *Z, W,* and thinking the thought that two sensations are similar in respect *R* if and only if they are similar to *Z, W',* then we are involved in a regress to infinity.)

The difficulty with the similitude theory of reference is the same as the difficulty with the 'causal chain of the appropriate kind' theory that we mentioned earlier. If I just say, 'The word "horse" refers to objects which have the property whose occur-

rence causes me on certain occasions to produce the utterance "there is a horse in front of me" ', then one difficulty is that there are too many such properties. For example, let H-A (for 'Horse Appearance') be that property of total perceptual situations which elicits the response 'there is a horse in front of me' from a competent normal speaker of English. Then the property H-A is present when I say 'there is a horse in front of me' (even when I am experiencing an illusion), but 'horse' does not refer to situations with that property, but rather to certain animals. The presence of an animal with the property of belonging to a particular natural kind and the presence of a perceptual situation with the property H-A are *both* connected to my utterance 'There is a horse in front of me' by causal chains. In fact, the occurrence of horses in the Stone Age is connected with my utterance 'There is a horse in front of me' by a causal chain. Just as there are *too many* similarities for reference to be merely a matter of similarities, so there are *too many* causal chains for reference to be merely a matter of causal chains.

On the other hand, if I say 'the word "horse" refers to objects which have a property which is connected with my production of the utterance "There is a horse in front of me" on certain occasions by *a causal chain of the appropriate type'*, then I have the problem that, if I am able to specify what *is* the appropriate type of causal chain, I must *already* be able to refer to the kinds of things and properties that make up that kind of causal chain. But how did I get to be able to do this?

The conclusion is not that there are *no* terms which have the logic ascribed by the similitude theory, any more than the conclusion is that there are *no* terms which refer to things which are connected to us by particular kinds of causal chains. The conclusion is simply that neither similitude nor causal connection can be the only, or the fundamental, mechanism of reference.

Wittgenstein on 'following a rule'

Consider the example I mentioned in passing, of the man who attempts to specify the respect R (the respect in which sensations must be similar to X if they are to be correctly classified as E) by saying or thinking that two things are similar in the respect R just in case they are similar in just the way Z, W are similar. This

fails, of course, because any two things *Z, W* are themselves similar in more than one way (in fact, in infinitely many ways). Trying to specifiy a similarity relation by giving finitely many examples is like trying to specify a function on the natural numbers by giving its first 1,000 (or 1,000,000) values: there are always infinitely many functions which agree with any given table on any finite set of values, but which diverge on values not listed in the table.

This is connected with another point that Wittgenstein makes in *Philosophical Investigations* and that was mentioned at the end of Chapter 1. Whatever introspectible signs or 'presentations' I may be able to call up in connection with a concept cannot specify or constitute the *content* of the concept. Wittgenstein makes this point in a famous section which concerns 'following a rule' – say, the rule 'add one'. Even if two species in two possible worlds (I state the argument in *most* un-Wittgensteinian terminology!) have the same mental signs in connection with the verbal formula 'add one', it is still possible that their *practice* might diverge; and it is the practice that fixes the interpretation: signs do not interpret themselves, as we saw. Even if someone pictures the relation '*A* is the successor of *B*' (i.e. $A = B + 1$) just as we do and has agreed with us on some large finite set of cases (e.g. that 2 is the successor of 1, 3 is the successor of 2, . . ., 999,978 is the successor of 999,977), still he may have a divergent interpretation of 'successor' which will only reveal itself in some future cases. (Even if he agrees with us in his 'theory' – i.e. what he *says* about 'successor of'; he may have a divergent interpretation of the whole theory, as the Skolem–Löwenheim Theorem shows.)

This has immediate relevance to philosophy of mathematics, as well as to philosophy of language. First of all, there is the question of *finitism:* human practice, actual and potential, extends only finitely far. Even if we say we can, we cannot 'go on counting forever'. If there are *possible divergent extensions of our practice, then there are possible divergent interpretations of even the natural number sequence* – our practice, or our mental representations, etc., do not single out a unique 'standard model' of the natural number sequence. We are tempted to think they do because we easily shift from 'we could go on counting' to 'an ideal machine could go on counting' (or, 'an ideal *mind*

could go on counting'); but talk of ideal machines (or minds) is very different from talk of *actual* machines and persons. Talk of what an ideal machine could do is talk *within* mathematics, it cannot fix the interpretation *of* mathematics.

In the same way, Wittgenstein holds that talk of 'similarity' and 'the same sensation' or 'the same experience' is talk *within* psychological theory; it cannot fix the interpretation *of* psychological theory. *That,* the interpretation of psychological theory and terminology, is fixed by our actual practice, our actual standards of correctness and incorrectness.

In *Ways of Worldmaking*[8] Nelson Goodman makes a closely related point: it is futile to try to have a notion of what the perceptual facts 'really are' independently of how we conceptualize them, of the descriptions that we give of them and that seem right to us. Thus, after discussing a finding by the psychologist Kolers that a disproportionate number of engineers and physicians are unable to see apparent motion at all, that is 'motion' produced by lights which successively flash at different positions, Goodman comments (p. 92):

> Yet if an observer reports that he sees two distinct flashes, even at distances and intervals so short that most observers see one moving spot, perhaps he means that he sees the two as we might say we see a swarm of molecules when we look at a chair, or as we do when we say we see a round table top even when we look at it from an oblique angle. Since an observer can become adept at distinguishing apparent from real motion, he may take the appearance of motion as a sign that there are two flashes, as we take the oval appearance of the table top as a sign that it is round; and in both cases the signs may be or become so transparent that we look through them to physical events and objects. When the observer visually determines that what is before him is what we agree is before him, we can hardly charge him with an error in visual perception. Shall we say, rather, that he misunderstands the instruction, which is presumably just to tell what he sees? Then how, without prejudicing the outcome, can we so reframe the instruction as to

[8] Published by Hackett, 1978.

> prevent such a 'misunderstanding'? Asking him to make no use of prior experience and to avoid all conceptualization will obviously leave him speechless; for to talk at all he must use words.

Grasp of 'Forms' and empirical association

A Platonist or Neo-Platonist of an antique vintage would have dealt with this issue in a much simpler way. Such a philosopher would have said that when we attend to a particular sensation we also perceive a Universal or a Form, i.e. the mind has the ability to grasp properties in themselves, and not just to attend to instances of those properties. Such a philosopher would say it is the Nominalism of Wittgenstein and Goodman, their refusal to have any truck with Forms and with the direct grasp of Forms, that makes it seem to them that there is any problem with the similitude theory.

While just positing a mysterious power of 'grasping Forms' is hardly a solution, it might seem that an analogue of this power *is* available to us. Properties of things do enter into causal *explanations;* when I have a sensation and it elicits the response 'this is a sensation of red', my response is partly caused by the fact that the sensation had a *property.* True, some philosophers are so nominalistic that they would deny the existence of such entities as 'properties' altogether; but science itself does not hesitate to talk freely of properties. Can we not say that, when Wittgenstein's privateer (the man who wanted to invent a private language) attended to *X* and said '*E*' then what *caused* the response '*E*' was a causal interaction involving a certain *property,* and *that* property (whatever it was) is the relevant 'similarity' that other sensations must have to *X* to be correctly classified as *E?*

The observation that talk of 'properties' is perfectly scientifically legitimate is correct; but this does not help rehabilitate Platonism. We interact with properties only by interacting with their *instances;* and these instances always are instances of *many* properties at the same time. There is no such thing as just interacting with a property 'in itself'. Talk of the properties causally associated with a sensation cannot do the work that the notion of the (unique) Form of the sensation did in Platonistic philosophy.

To spell this out: when I have a sensation of blue, I have a sensation of *blue,* and I also have a sensation with the complex property of being such as to be classed by me at that instance under that particular verbal label. Merely attending to *this* sensation does not constitute 'grasping' *one* of these properties. To pick out the property associated in just *one* of these ways with my sensation or with the verbal label is our old friend, the problem of the Causal Chain of the Appropriate Type again.

To see this, observe, first of all, that when my total perceptual experience elicits the response 'I am having the sensation of blue', I am not always *right.* I myself have had the experience of referring to 'the man in the blue sweater' two or three times before someone pointed out that the sweater was *green.* I don't mean the sweater *looked* blue; I realized that I had been misdescribing the sweater the instant the other person spoke. (I don't often have occasion to say 'I am having the sensation of blue', but if I did, then in such a case I would probably have said it two or three times until someone – wondering, perhaps, why I would have the sensation of blue when I was looking at something that was obviously green – queried me, whereupon I would have taken back my previous phenomenal report.) This already shows that the property of eliciting the report 'I am having the sensation of blue', or whatever, is *not the same property* as the property of being a sensation of blue, or a sensation of whatever the relevant quality might be.

Philosophers often refer to such a case as a 'slip of the tongue'. This seems to me to be an unfortunate terminology. The word 'green' might have been on my lips, and I might have found myself, frustratingly, *saying* 'blue'. *That* would have been a slip of the tongue. But in the case I described I didn't even notice I was misdescribing until someone questioned my report (and might never have noticed otherwise).

Another explanation which is suggested is that when I said 'blue' I *meant* green. By now it should be clear that when we say things we don't go around 'meaning' things in the sense of holding meanings in mind. To say I 'meant' green is just to say that I instantly accepted the correction (and felt funny when I realized the way I had been speaking). This is just to repeat what happened, not to explain it.

Whatever the explanation may be (perhaps some slip-up in the

verbal processing unit of my brain), the point is that, just as the property A-H described a few pages back will elicit the report 'There is a horse in front of me' even on occasions when no horse is present in the environment, so there is a complex property of my total mind-set which will elicit 'I am having a sensation of blue', when I am not having the sensation of blue (or, anyway, would deny that I was if I were queried). No mechanism of empirical association is perfect. If we decide to stipulate that I am having a sensation of blue whenever I am having a sensation which *elicits* that report (or which elicits that report and is such that the report does not seem 'wrong' to me on second thought), then on folk psychological theory, and perhaps on scientific psychological theory as well, there could be occasions when it will be true that I am having a sensation of blue by *this* criterion although, for one of a variety of reasons, the quality of the sensation is not blue. Moreover, as Wittgenstein puts it, on such a criterion, *whatever seems right to me is going to be right* – i.e. the distinction between making a report of my sensation that really is correct and making a report that seems to me to be correct will have been abandoned. Perhaps we *should* abandon or at least qualify it; perhaps, as Goodman seems to be suggesting, the question of whether one is 'really' having the kind of sensation one thinks one is makes no sense, apart from special cases, such as the case in which one would take the report back if queried; but to abandon this distinction is not a possible move for a metaphysical realist, for the sharp distinction between what really is the case and what one judges to be the case is precisely what constitutes metaphysical realism.

Could one always be wrong about the quality of one's past sensations?

Another way to bring out what is involved is to consider the question: 'Could one always be wrong about one's past sensations?' On the similitude theory, the answer is clearly 'yes'. For according to that theory, my previous sensations either are or aren't similar to the sensations I *now* describe by the various verbal labels 'sensation of red', 'pain', etc., and whether they are or aren't is a totally different question from whether I *then* classified them under those same verbal labels. Perhaps the world is

such that what we call a 'sensation of red' at an even numbered minute from the beginning of the Christian Era is actually similar in quality to what we call a 'sensation of green' at an odd-numbered minute, but our memory always deceives us in such a way that we never notice. Then the sensation I classified under the verbal label 'sensation of red' one minute ago would *not* be similar to the sensation I now classify under that same label.

There is something very odd about this alleged possibility, however. For one thing, the sense in which 'I would never notice' is very strong: if I treat my 'sensations of red' at different times as reliable signs of the various correlated physical occurrences (such as fire, the signal to stop, etc.) then I will be successful in all my actions. The 'wrong' similarity class (the class that lumps together the sensations I *call* sensations of red, in spite of the fact that they are not 'really' all of the same 'quality') would be the one that I had *better* use in connection with my problem-solving activities. But then is it really the *wrong* similarity class?

If we don't suppose that the notion of similarity is self-interpreting, then this case could be redescribed as a case in which the relation called 'similarity' by the external observer who is telling us about the case simply differs from the relation called 'similarity' by *us*. If we take this view, then the hypothesis that we are 'really' wrong about our past sensations collapses: from an *internalist* point of view there is no intelligible notion of sensations at different times being 'similar' apart from our standards of rational acceptability.

The correspondence theory of truth again

By now the reader may be convinced that the similitude theory of reference is thoroughly dead. But why should we conclude that the correspondence theory of truth must be given up? Even if the notion of a 'similarity' between our concepts and what they refer to doesn't work, couldn't there be some kind of an abstract isomorphism, or, if not literally an isomorphism, some kind of abstract *mapping* of concepts onto things in the (mind-independent) world? Couldn't truth be defined in terms of such an isomorphism or mapping?

The trouble with this suggestion is not that correspondences between words or concepts and other entities don't exist, but

that *too many* correspondences exist. To pick out just *one* correspondence between words or mental signs and mind-independent things we would have already to have referential access to the mind-independent things. You can't single out a correspondence between two things by just squeezing *one* of them hard (or doing anything else to just one of them); you cannot single out a correspondence between our concepts and the supposed noumenal objects without access to the noumenal objects.

One way to see this is the following. Sometimes incompatible theories can actually be intertranslatable. For example, if Newtonian physics were true, then every single physical event could be described in two ways: in terms of particles acting at a distance, across empty space (which is how Newton described gravitation as acting), or in terms of particles acting on fields which act on other fields (or other parts of the same field), which finally act 'locally' on other particles. For example, the Maxwell equations, which describe the behavior of the electro-magnetic field, are mathematically equivalent to a theory in which there are only action-at-a-distance forces between particles, attracting and repelling according to the inverse square law, travelling not instantaneously but rather at the speed of light ('retarded potentials'). The Maxwell field theory and the retarded potential theory are incompatible from a metaphysical point of view, since either there are or there aren't causal agencies (the 'fields') which mediate the action of separated particles on each other (a realist would say). But the two theories are mathematically intertranslatable. So if there is a 'correspondence' to the noumenal things which makes one of them true, then one can define another correspondence which makes the other theory true. If all it takes to make a theory true is abstract correspondence (never mind which), then incompatible theories can be true.

To an internalist this is not objectionable: why should there not sometimes be equally coherent but incompatible conceptual schemes which fit our experiential beliefs equally well? If truth is not (unique) correspondence then the possibility of a certain pluralism is opened up. But the motive of the metaphysical realist is to save the notion of the God's Eye Point of View, i.e. the One True Theory.

Not only may there be correspondence between objects and

(what we take to be) incompatible theories (i.e. *the same objects* can be what logicians call a 'model' for incompatible theories), but even if we fix the theory *and* fix the objects there are (if the number of objects is infinite) infinitely many *different* ways in which the same objects can be used to make a model for a given theory. This simply states in mathematical language the intuitive fact that to single out a correspondence between two domains one needs some independent access to both domains.

What we have is the demise of a theory that lasted for over two thousand years. That it persisted so long and in so many forms in spite of the internal contradictions and obscurities which were present from the beginning testifies to the naturalness and the strength of the desire for a God's Eye View. Kant, who first taught us that this desire is unfulfillable, thought that it was nonetheless built into our rational nature itself (he suggested sublimating this 'totalizing' impulse in the project of trying to realize 'the highest good in the world' by reconciling the moral and empirical orders in a perfected system of social institutions and individual relationships). The continued presence of this natural but unfulfillable impulse is, perhaps, a deep cause of the false monisms and false dualisms which proliferate in our culture; be this as it may, we are left without the God's Eye View.