1

Brains in a vat

An ant is crawling on a patch of sand. As it crawls, it traces a line in the sand. By pure chance the line that it traces curves and recrosses itself in such a way that it ends up looking like a recognizable caricature of Winston Churchill. Has the ant traced a picture of Winston Churchill, a picture that *depicts* Churchill?

Most people would say, on a little reflection, that it has not. The ant, after all, has never seen Churchill, or even a picture of Churchill, and it had no intention of depicting Churchill. It simply traced a line (and even *that* was unintentional), a line that *we* can 'see as' a picture of Churchill.

We can express this by saying that the line is not 'in itself' a representation[1] of anything rather than anything else. Similarity (of a certain very complicated sort) to the features of Winston Churchill is not sufficient to make something represent or refer to Churchill. Nor is it necessary: in our community the printed shape 'Winston Churchill', the spoken words 'Winston Churchill', and many other things are used to represent Churchill (though not pictorially), while not having the sort of similarity

[1] In this book the terms 'representation' and 'reference' always refer to a relation between a word (or other sort of sign, symbol, or representation) and something that actually exists (i.e. not just an 'object of thought'). There is a sense of 'refer' in which I can 'refer' to what does not exist; this is not the sense in which 'refer' is used here. An older word for what I call 'representation' or 'reference' is *denotation*.

Secondly, I follow the custom of modern logicians and use 'exist' to mean 'exist in the past, present, or future'. Thus Winston Churchill 'exists', and we can 'refer to' or 'represent' Winston Churchill, even though he is no longer alive.

to Churchill that a picture – even a line drawing – has. If *similarity* is not necessary or sufficient to make something represent something else, how can *anything* be necessary or sufficient for this purpose? How on earth can one thing represent (or 'stand for', etc.) a different thing?

The answer may seem easy. Suppose the ant had seen Winston Churchill, and suppose that it had the intelligence and skill to draw a picture of him. Suppose it produced the caricature *intentionally*. Then the line would have represented Churchill.

On the other hand, suppose the line had the shape WINSTON CHURCHILL. And suppose this was just accident (ignoring the improbability involved). Then the 'printed shape' WINSTON CHURCHILL would *not* have represented Churchill, although that printed shape does represent Churchill when it occurs in almost any book today.

So it may seem that what is necessary for representation, or what is mainly necessary for representation, is *intention*.

But to have the intention that *anything*, even private language (even the words 'Winston Churchill' spoken in my mind and not out loud), should *represent* Churchill, I must have been able to *think about* Churchill in the first place. If lines in the sand, noises, etc., cannot 'in themselves' represent anything, then how is it that thought forms can 'in themselves' represent anything? Or can they? How can thought reach out and 'grasp' what is external?

Some philosophers have, in the past, leaped from this sort of consideration to what they take to be a proof that the mind is *essentially non-physical in nature*. The argument is simple; what we said about the ant's curve applies to any physical object. No physical object can, in itself, refer to one thing rather than to another; nevertheless, *thoughts in the mind* obviously do succeed in referring to one thing rather than another. So thoughts (and hence the mind) are of an essentially different nature than physical objects. Thoughts have the characteristic of *intentionality* – they can refer to something else; nothing physical has 'intentionality', save as that intentionality is derivative from some employment of that physical thing by a mind. Or so it is claimed. This is too quick; just postulating mysterious powers of mind solves nothing. But the problem is very real. How is intentionality, reference, possible?

Magical theories of reference

We saw that the ant's 'picture' has no necessary connection with Winston Churchill. The mere fact that the 'picture' bears a 'resemblance' to Churchill does not make it into a real picture, nor does it make it a representation of Churchill. Unless the ant is an intelligent ant (which it isn't) and knows about Churchill (which it doesn't), the curve it traced is not a picture or even a representation of anything. Some primitive people believe that some representations (in particular, *names*) have a necessary connection with their bearers; that to know the 'true name' of someone or something gives one power over it. This power comes from the *magical connection* between the name and the bearer of the name; once one realizes that a name *only* has a contextual, contingent, conventional connection with its bearer, it is hard to see why knowledge of the name should have any mystical significance.

What is important to realize is that what goes for physical pictures also goes for mental images, and for mental representations in general; mental representations no more have a necessary connection with what they represent than physical representations do. The contrary supposition is a survival of magical thinking.

Perhaps the point is easiest to grasp in the case of mental *images*. (Perhaps the first philosopher to grasp the enormous significance of this point, even if he was not the first to actually make it, was Wittgenstein.) Suppose there is a planet somewhere on which human beings have evolved (or been deposited by alien spacemen, or what have you). Suppose these humans, although otherwise like us, have never seen *trees*. Suppose they have never imagined trees (perhaps vegetable life exists on their planet only in the form of molds). Suppose one day a picture of a tree is accidentally dropped on their planet by a spaceship which passes on without having other contact with them. Imagine them puzzling over the picture. What in the world is this? All sorts of speculations occur to them: a building, a canopy, even an animal of some kind. But suppose they never come close to the truth.

For *us* the picture is a representation of a tree. For these humans the picture only represents a strange object, nature and function unknown. Suppose one of them has a mental image

which is exactly like one of my mental images of a tree as a result of having seen the picture. His mental image is not a *representation of a tree*. It is only a representation of the strange object (whatever it is) that the mysterious picture represents.

Still, someone might argue that the mental image is *in fact* a representation of a tree, if only because the picture which caused this mental image was itself a representation of a tree to begin with. There is a causal chain from actual trees to the mental image even if it is a very strange one.

But even this causal chain can be imagined absent. Suppose the 'picture of the tree' that the spaceship dropped was not really a picture of a tree, but the accidental result of some spilled paints. Even if it looked exactly like a picture of a tree, it was, in truth, no more a picture of a tree than the ant's 'caricature' of Churchill was a picture of Churchill. We can even imagine that the spaceship which dropped the 'picture' came from a planet which knew nothing of trees. Then the humans would still have mental images qualitatively identical with my image of a tree, but they would not be images which represented a tree any more than anything else.

The same thing is true of *words*. A discourse on paper might seem to be a perfect description of trees, but if it was produced by monkeys randomly hitting keys on a typewriter for millions of years, then the words do not refer to anything. If there were a person who memorized those words and said them in his mind without understanding them, then they would not refer to anything when thought in the mind, either.

Imagine the person who is saying those words in his mind has been hypnotized. Suppose the words are in Japanese, and the person has been told that he understands Japanese. Suppose that as he thinks those words he has a 'feeling of understanding'. (Although if someone broke into his train of thought and asked him what the words he was thinking *meant*, he would discover he couldn't say.) Perhaps the illusion would be so perfect that the person could even fool a Japanese telepath! But if he couldn't use the words in the right contexts, answer questions about what he 'thought', etc., then he didn't understand them.

By combining these science fiction stories I have been telling, we can contrive a case in which someone thinks words which are in fact a description of trees in some language *and* simultane-

ously has appropriate mental images, but *neither* understands the words *nor* knows what a tree is. We can even imagine that the mental images were caused by paint-spills (although the person has been hypnotized to think that they are images of something appropriate to his thought – only, if he were asked, he wouldn't be able to say of what). And we can imagine that the language the person is thinking in is one neither the hypnotist nor the person hypnotized has ever heard of – perhaps it is just coincidence that these 'nonsense sentences', as the hypnotist supposes them to be, are a description of trees in Japanese. In short, everything passing before the person's mind might be qualitatively identical with what was passing through the mind of a Japanese speaker who was *really* thinking about trees – but none of it would refer to trees.

All of this is really impossible, of course, in the way that it is really impossible that monkeys should by chance type out a copy of *Hamlet*. That is to say that the probabilities against it are so high as to mean it will never really happen (we think). But is is not logically impossible, or even physically impossible. It *could* happen (compatibly with physical law and, perhaps, compatibly with actual conditions in the universe, if there are lots of intelligent beings on other planets). And if it did happen, it would be a striking demonstration of an important conceptual truth; that even a large and complex system of representations, both verbal and visual, still does not have an *intrinsic,* built-in, magical connection with what it represents – a connection independent of how it was caused and what the dispositions of the speaker or thinker are. And this is true whether the system of representations (words and images, in the case of the example) is physically realized – the words are written or spoken, and the pictures are physical pictures – or only realized in the mind. Thought words and mental pictures do not *intrinsically* represent what they are about.

The case of the brains in a vat

Here is a science fiction possibility discussed by philosophers: imagine that a human being (you can imagine this to be yourself) has been subjected to an operation by an evil scientist. The person's brain (your brain) has been removed from the body and

placed in a vat of nutrients which keeps the brain alive. The nerve endings have been connected to a super-scientific computer which causes the person whose brain it is to have the illusion that everything is perfectly normal. There seem to be people, objects, the sky, etc; but really all the person (you) is experiencing is the result of electronic impulses travelling from the computer to the nerve endings. The computer is so clever that if the person tries to raise his hand, the feedback from the computer will cause him to 'see' and 'feel' the hand being raised. Moreover, by varying the program, the evil scientist can cause the victim to 'experience' (or hallucinate) any situation or environment the evil scientist wishes. He can also obliterate the memory of the brain operation, so that the victim will seem to himself to have always been in this environment. It can even seem to the victim that he is sitting and reading these very words about the amusing but quite absurd supposition that there is an evil scientist who removes people's brains from their bodies and places them in a vat of nutrients which keep the brains alive. The nerve endings are supposed to be connected to a super-scientific computer which causes the person whose brain it is to have the illusion that . . .

When this sort of possibility is mentioned in a lecture on the Theory of Knowledge, the purpose, of course, is to raise the classical problem of scepticism with respect to the external world in a modern way. (*How do you know you aren't in this predicament?*) But this predicament is also a useful device for raising issues about the mind/world relationship.

Instead of having just one brain in a vat, we could imagine that all human beings (perhaps all sentient beings) are brains in a vat (or nervous systems in a vat in case some beings with just a minimal nervous system already count as 'sentient'). Of course, the evil scientist would have to be outside – or would he? Perhaps there is no evil scientist, perhaps (though this is absurd) the universe just happens to consist of automatic machinery tending a vat full of brains and nervous systems.

This time let us suppose that the automatic machinery is programmed to give us all a *collective* hallucination, rather than a number of separate unrelated hallucinations. Thus, when I seem to myself to be talking to you, you seem to yourself to be hearing my words. Of course, it is not the case that my words actually

reach your ears – for you don't have (real) ears, nor do I have a real mouth and tongue. Rather, when I produce my words, what happens is that the efferent impulses travel from my brain to the computer, which both causes me to 'hear' my own voice uttering those words and 'feel' my tongue moving, etc., and causes you to 'hear' my words, 'see' me speaking, etc. In this case, we are, in a sense, actually in communication. I am not mistaken about your real existence (only about the existence of your body and the 'external world', apart from brains). From a certain point of view, it doesn't even matter that 'the whole world' is a collective hallucination; for you do, after all, really hear my words when I speak to you, even if the mechanism isn't what we suppose it to be. (Of course, if we were two lovers making love, rather than just two people carrying on a conversation, then the suggestion that it was just two brains in a vat might be disturbing.)

I want now to ask a question which will seem very silly and obvious (at least to some people, including some very sophisticated philosophers), but which will take us to real philosophical depths rather quickly. Suppose this whole story were actually true. Could we, if we were brains in a vat in this way, *say* or *think* that we were?

I am going to argue that the answer is 'No, we couldn't.' In fact, I am going to argue that the supposition that we are actually brains in a vat, although it violates no physical law, and is perfectly consistent with everything we have experienced, cannot possibly be true. *It cannot possibly be true,* because it is, in a certain way, self-refuting.

The argument I am going to present is an unusual one, and it took me several years to convince myself that it is really right. But it is a correct argument. What makes it seem so strange is that it is connected with some of the very deepest issues in philosophy. (It first occurred to me when I was thinking about a theorem in modern logic, the 'Skolem–Löwenheim Theorem', and I suddenly saw a connection between this theorem and some arguments in Wittgenstein's *Philosophical Investigations*.)

A 'self-refuting supposition' is one whose truth implies its own falsity. For example, consider the thesis that *all general statements are false*. This is a general statement. So if it is true, then it must be false. Hence, it is false. Sometimes a thesis is called 'self-refuting' if it is *the supposition that the thesis is entertained*

or enunciated that implies its falsity. For example, 'I do not exist' is self-refuting if thought by *me* (for any '*me*'). So one can be certain that one oneself exists, if one thinks about it (as Descartes argued).

What I shall show is that the supposition that we are brains in a vat has just this property. If we can consider whether it is true or false, then it is not true (I shall show). Hence it is not true.

Before I give the argument, let us consider why it seems so strange that such an argument can be given (at least to philosophers who subscribe to a 'copy' conception of truth). We conceded that it is compatible with physical law that there should be a world in which all sentient beings are brains in a vat. As philosophers say, there is a 'possible world' in which all sentient beings are brains in a vat. (This 'possible world' talk makes it sound as if there is a *place* where any absurd supposition is true, which is why it can be very misleading in philosophy.) The humans in that possible world have exactly the same experiences that *we* do. They think the same thoughts we do (at least, the same words, images, thought-forms, etc., go through their minds). Yet, I am claiming that there is an argument we can give that shows we are not brains in a vat. How can there be? And why couldn't the people in the possible world who really *are* brains in a vat give it too?

The answer is going to be (basically) this: although the people in that possible world can think and 'say' any words we can think and say, they cannot (I claim) *refer* to what we can refer to. In particular, they cannot think or say that they are brains in a vat (*even by thinking 'we are brains in a vat'*).

Turing's test

Suppose someone succeeds in inventing a computer which can actually carry on an intelligent conversation with one (on as many subjects as an intelligent person might). How can one decide if the computer is 'conscious'?

The British logician Alan Turing proposed the following test:[2] let someone carry on a conversation with the computer and a conversation with a person whom he does not know. If he can-

[2] A. M. Turing, 'Computing Machinery and Intelligence', *Mind* (1950), reprinted in A. R. Anderson (ed.), *Minds and Machines*.

not tell which is the computer and which is the human being, then (assume the test to be repeated a sufficient number of times with different interlocutors) the computer is conscious. In short, a computing machine is conscious if it can pass the 'Turing Test'. (The conversations are not to be carried on face to face, of course, since the interlocutor is not to know the visual appearance of either of his two conversational partners. Nor is voice to be used, since the mechanical voice might simply sound different from a human voice. Imagine, rather, that the conversations are all carried on via electric typewriter. The interlocutor types in his statements, questions, etc., and the two partners – the machine and the person – respond via the electric keyboard. Also, the machine may *lie* – asked 'Are you a machine', it might reply, 'No, I'm an assistant in the lab here.')

The idea that this test is really a definitive test of consciousness has been criticized by a number of authors (who are by no means hostile in principle to the idea that a machine might be conscious). But this is not our topic at this time. I wish to use the general idea of the Turing test, the general idea of a *dialogic test of competence,* for a different purpose, the purpose of exploring the notion of *reference*.

Imagine a situation in which the problem is not to determine if the partner is really a person or a machine, but is rather to determine if the partner uses the words to refer as we do. The obvious test is, again, to carry on a conversation, and, if no problems arise, if the partner 'passes' in the sense of being indistinguishable from someone who is certified in advance to be speaking the same language, referring to the usual sorts of objects, etc., to conclude that the partner does refer to objects as we do. When the purpose of the Turing test is as just described, that is, to determine the existence of (shared) reference, I shall refer to the test as the *Turing Test for Reference*. And, just as philosophers have discussed the question whether the original Turing test is a *definitive* test for consciousness, i.e. the question of whether a machine which 'passes' the test not just once but regularly is *necessarily* conscious, so, in the same way, I wish to discuss the question of whether the Turing Test for Reference just suggested is a definitive test for shared reference.

The answer will turn out to be 'No'. The Turing Test for Reference is not definitive. It is certainly an excellent test in practice;

but it is not logically impossible (though it is certainly highly improbable) that someone could pass the Turing Test for Reference and not be referring to anything. It follows from this, as we shall see, that we can extend our observation that words (and whole texts and discourses) do not have a necessary connection to their referents. Even if we consider not words by themselves but rules deciding what words may appropriately be produced in certain contexts – even if we consider, in computer jargon, *programs for using words* – unless those programs themselves *refer to something extra-linguistic* there is still no determinate reference that those words possess. This will be a crucial step in the process of reaching the conclusion that the Brain-in-a-Vat Worlders cannot refer to anything external at all (and hence cannot say *that* they are Brain-in-a-Vat Worlders).

Suppose, for example, that I am in the Turing situation (playing the 'Imitation Game', in Turing's terminology) and my partner is actually a machine. Suppose this machine is able to win the game ('passes' the test). Imagine the machine to be programmed to produce beautiful responses in English to statements, questions, remarks, etc. in English, but that it has no sense organs (other than the hookup to my electric typewriter), and no motor organs (other than the electric typewriter). (As far as I can make out, Turing does not assume that the possession of either sense organs or motor organs is necessary for consciousness or intelligence.) Assume that not only does the machine lack electronic eyes and ears, etc., but that there are no provisions in the machine's program, the program for playing the Imitation Game, for incorporating inputs from such sense organs, or for controlling a body. What should we say about such a machine?

To me, it seems evident that we cannot and should not attribute reference to such a device. It is true that the machine can discourse beautifully about, say, the scenery in New England. But it could not recognize an apple tree or an apple, a mountain or a cow, a field or a steeple, if it were in front of one.

What we have is a device for producing sentences in response to sentences. But none of these sentences is at all connected to the real world. *If one coupled two of these machines and let them play the Imitation Game with each other, then they would*