LANGUAGE AND THE "COGNITIVE REVOLUTIONS"

A TEN-PART LECTURE SERIES

GIVEN BY

PROF. NOAM CHOMSKY

AT

THE UNIVERSITY OF GIRONA, CATALONIA

BETWEEN

23rd AND 27th NOVEMBER, 1992

UNOFFICIAL TRANSCRIPT

All information about the lecture series, including the audio files, can be found at https://www.catedraferratermora.cat/llicons/en/chomsky/

The video recordings can be found at https://www.youtube.com/watch?v=S3gFaNYluBQ/

Lecture #1

I WANT TO talk about what is called the "cognitive revolution" that began in the 1950's. When I suggested the topic, I also suggested that quotes be put around the words 'cognitive revolution', suggesting a certain skepticism about its revolutionary character. The skepticism has two aspects to it. One, I'm not convinced that it was as much of a change as many other people think. In many ways, it picked up and recovered ideas, even technical ideas, that are much older and that we can trace to what might be called more properly the first cognitive revolution of the 17th century. And it didn't always improve on those ideas. In some cases, I think there's been regression. So, it's not as revolutionary as it has been held to be. The earlier tradition had been entirely forgotten by the 1950's. Even in scholarship it was not known and understood, that is, scholars knew the books but they didn't understand what was in them. In my opinion, that's still more or less the case. I don't think the riches of the first cognitive revolution have yet been appreciated or understood, except in certain corners of scholarship.

The second aspect of skepticism is that the cognitive revolution has taken from the very beginning a rather dubious path, maybe a wrong turn, and that the directions in which it is proceeding should be seriously reassessed from their very origins. And here I would go back further. In the case of the study of language and mind back to Gottlob Frege, where I think already there were some intellectual moves made that are quite dubious and that have had questionable effects – in my view, negative effects – on the contemporary theory of reference and many other topics.

When I speak of the cognitive revolution, I don't mean to be referring to specific work in (say) the neurophysiology of vision, or the study of reasoning under complex conditions, and so on. A lot of that work is very respectable, scientific work. I'm thinking of the more reflective and considered aspects of the cognitive revolution, those that fall roughly within 'philosophy of mind' or so-called 'artificial intelligence' – parts that are concerned with the general nature of the issues rather than constructing an expert system that will solve some technical engineering problem. When I refer to the cognitive revolution with skepticism, it's at a general level, it's where it intersects with (or falls under, one might say) contemporary philosophy of mind and philosophy of language. Well, that's the skepticism, and I'll come back to trying to fill in the blanks.

The study of language and mind, as everyone knows, goes back several millennia, back to classical antiquity. It's often been assumed that these two inquiries – the inquiries into language and into mind – are intimately related, that language is a mirror of the mind, as Gottfried Leibniz put it. If that is the case, then the study of language should provide unique insight into human thought. That's often been thought to be the case over the last several thousand years. There have been repeated convergences between the more technical study of language and the more general study of mental events, actions, processes, and so on.

Now, this convergence has taken place in these two cognitive revolutions. It took place about 40 years ago at the origins of what is today commonly called the cognitive revolution, and contemporary linguistics developed as part of that cognitive revolution, and in fact has been a significant factor (maybe the major factor) in the development of cognitive science since modern origins. This same convergence took place during what I call the first cognitive revolution of the 17th century, which was part of the general scientific revolution of the period (the Galilean revolution).

The convergence took place in ways that are rather strikingly similar to the convergence of the 1950's in a number of respects. One respect was the stimulus to the scientific imagination that was given by automata, and in the 20th century that has, of course, been computers. In the 17th and 18th centuries, it was the remarkable automata that were constructed by skilled craftsmen, starting with extremely complicated clocks and reaching up to the creations of people like Jacques de Vaucanson (duck digesting food, etc.). In both the 17th and 18th centuries and in contemporary discussion, the apparent achievements of artifacts raised an obvious question, namely, whether humans are not simply more complex machines or artifacts. That was a topic of a very lively debate then as it is today.

The Cartesians, notoriously, offered a negative answer to this position. They said, no, humans are not more complex artifacts, although Descartes tried to show that a very substantial part of what humans do, including all the way up to perception and sensation, is just a complex watch or a complex machine. And the same is true, he argued, of the entire inorganic world, and organic world up to the level of (metaphorically speaking) humans below the neck. But he also argued that certain aspects of human intelligence lie beyond the scope of any conceivable artifact. He appealed specifically to language in this connection. That's where the convergence took place. Crucially, if you look at *Discourse on Method* and other Cartesian discussions, he appealed to normal aspects of language use as evidence of the kind of phenomenon that in principle could not be incorporated within an automaton,

even of the most complex and highly articulated kind. Specifically, he referred to a collection of properties that normal linguistic behaviour manifests. We can call them (though he didn't call them this) *the creative aspect of language use*. This is a collection of properties that includes the fact that normal language use (I don't mean poetic discourse but just ordinary interaction among people) is unbounded. People are always saying and hearing new things that have never been said before in the history of the human species, and this goes on constantly and unrecognizably. You have no way of knowing whether they're new or not because they all sound familiar, even if you've never heard them before. So, normal speech is unbounded. On the other hand, it is obviously not random. It's not just some device producing things with a random element in it. It's coherent. It's appropriate to situations, but it doesn't seem to be caused by situations. In fact, it appears to be completely uncaused. That's not to say that there aren't any influences, but it seems to be a paradigm example of the general matter of freedom of the will. If you had a complete description of your internal state and the surroundings that you're in, still, the Cartesians argue (and phenomenally they seem to be correct), you could choose to say something other than what is suggested, maybe even strongly pressured, by the internal state and the external environment. I could right now start talking about the weather in Boston or any other topic, or I could find topics that none of you would be able to even understand – what's happening in my family or something like that. All of this is always possible, and we know that it is possible. Even given my internal state completely described and my surroundings completely described, I can start doing anything.

This, of course, is a much more general problem, as the Cartesians emphasize. In their formulation, a machine is compelled to act in a certain way up to randomness. Up to randomness, a machine is compelled to act in a certain way by the arrangement of its parts and its stimulus situation. A human being, in contrast, is only incited or inclined to act in a certain way, and may choose not to. It may choose to act in a way contrary to its inclinations. You can choose to act, say, suicidally, and people sometimes do. That kind of phenomenon, which is manifested more clearly in the normal use of language, Descartes argued, lies beyond the bounds of any possible automaton. So, the fact that language is unbounded, stimulus-free, not determined by the internal state (though heavily influenced by it), appropriate to situations but not caused by it (apparently uncaused), non-random, evoking thoughts in others, etc – this collection of properties (the creative aspect of language use) is, Descartes argued, a kind of litmus test. Just as the litmus test is a test for acidity, these properties are a test for some property of the world that does not fall within mechanism.

Now, those arguments are not inconsiderable, and I don't think they change at all as we move from the complex artifacts that excited the imaginations of the Cartesians to the contemporary artifacts like computers. Although both are radically different in all sorts of ways, the difference between appropriateness to situations vs. caused by situations, appropriate use vs. random/arbitrary intrusion into deterministic systems seems to remain. The automata that we have are either deterministic or have elements of randomness, meaning they could be probabilistic in their behaviour. But none of that matches in the least the properties of ordinary human behaviour, so it appears. We could be completely wrong about this. Maybe we are misled by the facts. But these facts, as the Cartesians argued, seem as obvious to us as anything could be on immediate inspection. And, Descartes again argued, it would be absurd to deny that which appears obvious to us simply on the grounds that we do not now have intelligence enough to understand it – that would simply be irrational. Therefore, he argued that we have to take these things very seriously. I think we are still in this situation. That was the Cartesian response to the first question – whether people are simply more complex automata. Their argument was that everything is just a complex automaton, except certain aspects of human behaviour – language being a paradigm case.

Those issues arose again in the second cognitive revolution, and in the contemporary debate about the Turing test, and in all kinds of things Roger Penrose, for example, looks at in his recent book (*The Emperor's New Mind*), and the vast debate about John Searle's Chinese room in the philosophy of mind and artificial intelligence (AI) literature, and so on. All of this is a contemporary version of it, and in my view a case of regression, however. I think it is extremely and seriously misguided. As the question arose in the 17th century, it was a sensible question. Maybe the answers weren't sensible, but the question was sensible and the framework was sensible. I'll try to suggest later that the contemporary framework makes no sense whatsoever and is completely off-track, and, in my opinion, Turing in his original paper in 1950 already pointed out why it was off-track – that is, it's been following a line that he suggests people not follow. I'll come back to that.

The second similarity between the first and the second cognitive revolution is that in both cases there was great interest in computational theories of the mind. In fact, the great scientific achievement of Descartes, his major contribution to modern science, was the development of a computational theory of vision. He overthrew the prevailing neo-scholastic theories of vision, which had a very mystical character to them (kind of common-sensical but mystical). At the time, it was assumed that, say, if I see a cube rotating in space, then there's a cube rotating in space in my brain. Somehow the form of the thing out there by some mysterious process gets into my

brain, and that's what vision is – it's picking out the form of the object and duplicating it somehow. Descartes ridiculed this idea, properly, and not only tried to show why it was absurd but offered an alternative theory, which was a real scientific breakthrough. You can't use the theory in his form, but it led to the modern biology of cognitive processes.

In order to undermine this neo-scholastic view of vision, he suggested that we consider the case of a blind man with a stick who is tapping on a chair and getting a sequence of stimuli in his hand, and from this sequence he figures out that it's a chair that he's perceiving. Obviously, the image of the chair isn't getting into this brain. The only thing that's getting into his brain are some pressures against his fingers. There's a sequence of pressures on his fingers from which his mind is constructing the image of a chair somehow. Descartes argued that that's exactly what normal vision is. Given Cartesian physics, he had to assume that there's a solid connection between the retina and the object you see – that there's like a rigid rod that extends from your eye to the thing you see, and as you move your eye around, it's exactly like tapping on the chair with a stick, and it just happens that you're getting stimuli on your retina instead of your hand, but the picture is the same. He, therefore, argued that normal vision is just the computational interpretation by the brain of a sequence of pressures on the retina, which is exactly what the blind man is doing. That leads to a kind of computational theory of vision. It's the inner resources of the mind that determine what you see. He argued that (he didn't do the experiment but we could do the experiment now and see that his guess was right) if you take an infant who's never seen a geometrical figure and you present him with a triangle, what the child will see is a distorted Euclidean triangle (because of course what you've drawn is not a real triangle – two of the lines don't quite come together, one of them has got a curve, and so on). The child, in other words, will not perceive a perfect example of what it is (which is some crazy figure) but will see it as a distorted triangle, although the child doesn't have any experiences. The reason for this, Descartes argued, is that the mind operates on the principles of Euclidean geometry, and when a sequence of stimuli hits the retina, the child's mind creates the Euclidean abstract figure and that's what's seen, and you then notice that what's out there is a distorted version of it. He argued that this is generally true of perception.

That whole picture seems to be essentially correct, and it was a scientific breakthrough, and led the way to serious inquiry into the biology of vision and perception generally. That, I suppose, is one of Descartes's leading scientific contributions. It opened up modern physiology. Similar ideas have re-emerged in the 20th century cognitive revolution and have led to quite productive work in, strikingly, the very same areas that were explored in the 17th century – primarily

vision, and a few other sensory modalities, and language, and a little bit in other areas like conceptual development, reasoning, etc.

In the area of language, the Cartesian revolution did lead, at once, to efforts to apply this kind of computational point of view. It wasn't regarded as computational then but we would regard it now as computational. This computational point of view, which views a mechanical device from a certain abstract perspective, is the view of having the property of a computational device, that is, a collection of states and properties that the device could have. This account that I gave of the infant seeing the triangle could be re-described as a kind of software matter, if you like. That conception re-emerged somewhat reformulated in the 20th century. In the Cartesian period, it set off quite important studies of language, in fact, revolutionary studies, developing into what was then called 'rational' and 'philosophical' grammar, which just means 'scientific' in our terms, so that would be 'scientific linguistics'. It led to a conception of Universal Grammar (UG), i.e., the properties of language that are common to language generally and not to specific languages, which, of course, would be a core part of any scientific approach to studying language. It also led to the first real studies of the vernacular (like French), which was unusual at the time. The very fact that Descartes wrote in French was considered a real breakthrough, which has all sorts of political aspects to it as well. All of these were important achievements later forgotten, but reconstructed in many ways in the 20th century cognitive revolution.

The second cognitive revolution has indeed led to real advances in certain areas, strikingly, the old areas – vision, language, etc. – but it's not clear (to me at least) that it has led to any real progress in what you might consider a second level – the level of reflection about the nature of the disciplines that are concerned with what was traditionally called 'mental acts and faculties.' That's the question that I want to come back to as I proceed. Some of these questions are substantive and some are historical. Let me just start with a couple of words about the second cognitive revolution, the one that contemporary linguistics was part of and came out of, and then go back to a general look of the whole topic.

None of these things have a starting date, but George A. Miller (one of the leading figures in the birth (I'd call the rebirth) of cognitive psychology), in a recent retrospective talk on this, traced the cognitive revolution to a meeting that took place in Cambridge, Massachusetts. In 1956, there was a meeting of the Institute of Radio Engineers (IRE) (most of this work developed within an electrical engineering framework in the 1960's). At that meeting, there was a kind of unexpected convergence of different things, which, in retrospect, Miller argued, set off the

cognitive revolution, or at least gave it a kind of form. There were some papers in human information processing (human psychology) using new ideas like signal detection theory and information-theoretic ideas and so on. There was the first paper on generative grammar that had been given publicly – a paper of mine – which outlines some of the basic ideas that later became generative grammar. There was also an important paper on problem-solving and reasoning by Allen Newell and Herbert Simon. They gave their first exposition on their paper on a program for proving theorems in elementary logic, which was one of the things that set off contemporary AI. I should say that Newell and Simon didn't follow the path that became conventional, but this kind of set things off. It was considered a major stimulus to contemporary AI, which hadn't really been formulated yet at that time.

That was a collection of papers, and they came from different sources and proceeded in different ways, and there had not really been much communication among the people who had given them. But they had some shared ideas. One shared idea was a certain kind of shift in perspective from what we might call an 'externalist' point of view to an 'internalist' point of view. That is, the psychology of the time was externalist, in that it was concerned with what was outside the person. It was concerned with behaviour or the products of behaviour. So, linguistics was the study of products of behaviour, the study of texts, arrangement of words, structures of sounds, etc. That was true of both European and American structuralism (different in many respects but both externalist in this respect). Behaviourist psychology is a paradigm example of this – you're only interested in what's outside the mind. In fact, it's kind of a point of principle that you're not supposed to look at anything else. The shift that took place was from this externalist point of view to an internalist one in which what you're interested is in is precisely what's going on inside the mind-brain, where we think of the mind as just some set of properties, states, and processes of the brain, ultimately to be related to the brain sciences in some fashion. We use the term 'mind' here without any metaphysical implications.

The topic of inquiry shifts totally. Instead of the topic of inquiry being behaviour and the products of behaviour, the topic of inquiry is what's going on inside the mind. Behaviour and its products are just data, and not particularly privileged data. Data in itself is not good or bad. Data becomes significant when it becomes evidence, and evidence is a relational concept – evidence *for*. Data moves into the sciences when it becomes evidence for something, and that something is something about the nature of the world – in this case the nature of the mind. In my opinion, this shift is a shift from natural history (like rock collecting or something) to the beginnings of what might turn out to be science. It's a kind of shift from natural history to natural science, where behaviour and products of behaviour are simply

data, and not particularly privileged – useful if they're useful, otherwise throw them out because most of it is useless. If you could find evidence from electrical stimulation of the brain that could tell you something about language, that's just as good, maybe even better than evidence about the way people interpret sentences, or what they do with words, etc. Data is of no interest in itself.

Intellectually speaking, that was an enormous shift, and it was very controversial, and it remains very controversial today. If you take such criteria as government funding, government funding in the U.S. overwhelmingly goes to the externalist work – statistical analysis of texts, organization of data, etc. Very little of it goes to the internalist work, which is the only kind of work that even merits being talked about seriously. Anyhow, that's a personal view. But there was clearly a shift of perspective that was taking place at one or another level in of all this work. I don't want to exaggerate much because many people in the cognitive sciences regard this shift of perspective as dubious and wrong, but at least there were hints of it in all this work, and you can maybe see better in retrospect that a move was taking place in that direction. Again, in my opinion, that's a move from natural history, rather boring natural history, to a potential natural science.

A second set of shared ideas was in what you might call 'computational-representational' theories, i.e., looking at what the brain is doing as a kind of software problem. To look at it as a software problem is to take a certain abstract perspective toward the functioning of the machine. It's a perspective that sometimes makes sense and sometimes doesn't. Whenever you're studying some physical object, sometimes a particular abstract perspective makes sense, giving you insights, and sometimes it doesn't. That's what science is about. In this case, there was a kind of an intuitive shared feeling that viewing the brain as having hardware and software properties would be useful. Remember, this is totally abstract. Even in the case of a computer, when we distinguish between hardware and software, you can't really pull out a particular piece of it and say 'I'm hardware.' I mean, everything is just hardware. To say that a computer is implementing some software is to view it from a certain abstract perspective, which may or may not make sense. In the case of the brain, the questions may or may not make sense, much as looking at the planets in terms of rational mechanics in which you have mass points observing Newton's laws may or may not make sense. But it's a matter of discovery, not stipulation.

This point of view was liberating, as it had been in the 17th century. In the 17th century, they didn't talk about hardware and software, but I think it makes a lot of sense to reinterpret Descartes's overthrow of the neo-scholastic theory of vision in terms of the picture that I just presented – as adopting a computational-

representational point of view and separating the software aspects (the computational aspects which give you these idealized figures) from the hardware properties (like the sequence of stimuli, the finger, the brain, etc) – and in contemporary work, it's often looked at like that. If you look at the work in the David Marr school, that distinction is made quite explicit, and correctly, since it has led to a lot of progress – which is the only mark of correctness.

This is a double-edged sword, in my opinion. The move to a position where you look at mental activities as software problems was liberating. It was a move that should've been made. On the other hand, it can also be extremely misleading, and it has long become more misleading than helpful, particularly in the philosophy of mind. Those areas of the philosophy of mind that are kind of around AI, including inquiry into Turing tests, etc., have been seriously misled by the metaphors. I'll come back to that. Analogies, metaphors, and abstractions are fine, but you don't want to get misled by them, and that's never easy.

Let me now approach the question from an orthogonal point of view and ask how might one proceed to study humans altogether. There's one obvious idea – you should study them as being part of the natural world. This was traditionally seen to be controversial (for religious reasons), but in the post-Galilean period it should be possible to entertain the idea that you should study humans as part of the natural world. That doesn't mean that when you study humans, you'll find what you find when studying rocks, obviously. It just means that the method of approach should be as if they're part of the natural world, meaning that you should search for intelligible explanatory theories that give you some insight into what these objects are about, what they're up to, how they're constituted, etc., and you should hope that, ultimately, this inquiry will be integrated with other aspects of the natural sciences.

Typically, over the centuries, parts of the natural sciences proceed in isolation, and you don't know how to integrate them, but when you can integrate them then it's a big discovery. When biology was more or less incorporated in biochemistry about 40 years ago, that was a real breakthrough. When it became possible for physics to understand for the first time such things as why a solid object can exist (which was incomprehensible to 19th century physics), it was a real breakthrough. When it became possible to understand and incorporate within physics elementary properties of the world like states of matter, or the properties of solids, or the colour of sodium, or the character of the periodic table – in other words, the quantum-theoretic revolution – that was a real revolution. Nevertheless, quite often in normal science things just can't be integrated, because you don't know how to integrate

them, so they proceed separately, but with an eye to eventual integration. That's normal science. We get these miraculous moments when things get integrated, but that's not the norm by any means.

This approach that says let's study humans the way we study anything else, let's call it 'naturalistic'. By calling it 'naturalistic', I mean to try to focus attention on the character of work and to reasonable goals, and to abstract away from the question of success. Maybe it's completely unsuccessful, but that's another issue. So, I don't want to use the honorific term 'science' for it, but just *like* science.

Well, that's a kind of a common-sense idea. A naturalistic approach would claim that it has no burden of proof to meet at all – that it's self-justifying. Maybe there's some reason not to adopt the naturalistic point of view to humans, but that is what needs justification. The burden of proof is on anyone who questions this idea, or so a naturalistic picture would assume. We should agree that unless some argument is given (which hasn't yet been given), there's no reason not to study humans the way we would study rocks or bees or anything else – expecting to discover totally different things, of course (as when we study solids and liquids, we learn totally different things).

There are interesting questions as to how naturalistic inquiry ought to proceed – what are the criteria of rationality for science? What about the reality of theoretical entities? Does it make sense to claim that a mathematical object like a vector field has mass? But if one is interested in getting answers to these questions, and in not just harassing emerging disciplines, then the place to ask is where there might be a chance of getting an answer. That's the way you proceed if you're rational. In this case, that means physics and not psychology. You might find some answers in physics for obvious reasons – the depth of understanding and the degree of success is so qualitatively greater by orders of magnitude that there really are guides to inquiry into these questions. Whereas in the emerging disciplines, half the time you don't know what you're doing at all, so there's just nothing around to guide inquiry. If you were to ask these questions in physics in the Galilean period, you would've gotten all wrong answers. We know that. In fact, Galileo had a very hard time convincing anyone. He had a rather sad fate precisely because not enough was understood about these questions so that what he was trying to do could be intelligible. You can only get some insight where there are some advances. The attempts to raise these questions about fields like psychology and linguistics seems to me just a form of harassment. It's of no intellectual interest, though you can see why philosophers do it. Quantum physics is hard. If you want to raise questions about quantum physics, you must study hard, learn, think about things, etc. If you

want to raise questions about psychology, you can do it out the top of your head, because nothing much is known, and it makes life easier. But that's not a good reason to harass psychology, I don't think. Insofar as these general questions arise, I think we can dismiss them. Unless some special reason is given to show that psychology and linguistics have some special methodological problem that (say) physics doesn't have, general questions about induction, indeterminacy, etc., can be forgotten about. They all arise in physics, and if anything arises in physics, we can forget about it. If you want to get an answer, look over there where they have some hope of moving forward to some insight. There's no point in raising those questions here where so little is understood. So, I'll put those questions aside unless there's an argument.

Incidentally, this throws out an awful lot of contemporary philosophy, maybe unfairly, but I think fairly. A lot of it, in my view, is just harassment of emerging disciplines, hence we can put it aside. And the criterion that I want to use is this –

If some general methodological consideration can be shown to hold of language and psychology but not of chemistry and geology, then we'll consider it, but if it holds of chemistry and geology, we'll forget about it.

That's the working criterion that I'd like to suggest, based on the assumption that when we ask questions, we expect answers, not to harass. If we want answers, we look at the place where we might find them. That's the logic of the criterion I'm suggesting.

A naturalistic approach to humans, then, will put aside any general methodological issues that just arise in the course of rational inquiry, and will simply take for granted what is done in normal science (with all its uncertainties, equivocations, problems, etc.), always with an open mind, and will proceed to study humans in the way we study anything else – you try to find an intelligible explanatory theory that gives you some insight, opens up new paths of investigations, leads you into new kinds of empirical questions you hadn't thought of before, etc. Well, that's the naturalistic approach.

An alternative approach which rejects naturalism we might call 'dualistic'. A dualistic approach says that humans just aren't part of the natural world, and you must study them in quite a different way. Here we must be able to be more careful, because a naturalistic approach could lead to a certain kind of dualism. For example, Cartesian metaphysical dualism was completely naturalistic, that is, it was the outcome of a way of looking at human beings naturalistically, and it led to a conception that humans have some special property, just as acids have some special

property bases don't have. There's nothing dualistic about chemistry because acids are different from bases, and there's nothing non-naturalistic about Cartesian metaphysical dualism if it says that there is some special property that humans above the neck have, and that there's a litmus test for it, namely, things like the creative aspect of language use. It can be wrong, but it's not non-naturalistic.

I don't mean dualism and naturalism to be exclusive categories. There's a form of dualism, namely, metaphysical dualism (the traditional form) which could follow from a naturalistic approach and could even be right. The kind of pernicious and irrational dualism is a kind of epistemological dualism, not metaphysical dualism, the one that says that you're not allowed to approach humans by the same procedures of rational inquiry that you apply elsewhere. That may seem to be a crazy point of view, and I mean to suggest that it is indeed a crazy point of view, but I'll also suggest that it's the overwhelmingly dominant point of view in the philosophy of mind and the cognitive sciences. The view may be almost universal. I want to take a strong and controversial position on that; therefore, I'm putting it in the craziest possible form, and then I'll try to argue that what is done falls under this rubric.

Let's take traditional Cartesian metaphysical dualism. In outline, Descartes argued as follows. He had a conception of the physical world. The conception was what was called in those days 'mechanical philosophy'. 'Philosophy' was just a word for science. So 'mechanical philosophy' meant mechanics, and the mechanics was kind of common-sensical, a sort of contact mechanics – the crucial idea being that things can influence one and other if they're in contact. I can't move the moon by moving my arm because they're not in contact (it happens to be wrong, but it is common sense). That's common-sense contact mechanics, and Descartes gave a kind of sketchy account of how you might cover all of the phenomena of the inorganic world, the organic world, and most of human beings (except the things I mentioned) in terms of mechanical philosophy. Then he said that phenomena like the creative aspect of language use don't fall under this, hence we need a new principle. This was completely rational. Within his framework, the only way to set up a new principle was to introduce a new substance; therefore, he introduced a *res cogitans*, a thinking substance, which has other properties, and then from a naturalistic point of view you have two questions. First, what is the *res cogitans*? Nobody knows whether Descartes had an answer to this. If you look at his major work *Traité du monde*, there's only three volumes. There was a fourth volume which was, supposedly, devoted to the mind, and legend has it that he destroyed that volume after he heard what had happened to Galileo. Whether that's true or not I don't know. Maybe all the secrets are in there, like the Fermat's last theorem or something. In any event, there's nothing much around from the theory of mind, but

that would've been the problem – that if you've got this new substance then tell us what it is.

The second problem would be the standard unification problem that arises in the sciences altogether, which is to show how this theory relates to other theories. In the terminology of that day, it meant solving the interaction problem of showing how mind and body interact. The way in which the standard unification problem of normal science was stated was in terms of the two-substance theory of the structure of science. Remember, two substance theory of science was perfectly rational. Maybe completely wrong, but most theories have always been wrong. There's nothing irrational about it. Without doing serious violence to it, we can, I think, provide some understanding to its history by rethinking it in this form.

So, we have a naturalistic approach which claims to show that the whole range of phenomena of the world falls under contact mechanics. We identify some phenomena that don't fall into it. We invoke some new principle. Then we proceed to study that new principle by developing a theory of mind. And then we proceed to solve the unification problem, namely, to show how it falls in with other parts of the theory of the world.

It looks as if this is what Descartes was trying to do in *Principles of Philosophy* and *Traité du monde* and all that sort of stuff. Incidentally, these are the parts of Descartes that nobody ever reads. The parts of him that people read are things like *Discourse on Method*, which was like a research grant proposal, or *Meditations*, where he is trying to answer questions that philosophers raised about all this. But from Descartes's point of view, much as you can reconstruct, it seems that what was important was the science (like *Dioptrique*, etc.), and the other stuff was thinking about the nature of science. Since the science is now all outmoded, you don't study it (or, for that matter, nobody much studies Galileo either, because it is done differently and better), but that doesn't mean that it wasn't important to Descartes. What remains is the talk about the science, and since there's been no progress on that at all, you can make perfectly good sense to read *Discourse on Method*. But I think one should read it recognizing that for Descartes it was peripheral. If you want to understand the project, you have to look at it his way. When you reconstruct it his way, I think you find a very naturalistic project that ends up with metaphysical dualism as a serious proposal about the nature of the world, which can very well turn out to be right (like other serious proposals about the nature of the world). So, again, this is dualism but it's not non-naturalistic. It may be wrong, but it's naturalistic and within the spirit of science.

All of this has been ridiculed, especially in the modern period, as the idea that there is (as Gilbert Ryle put it) a ghost in the machine, and then you're supposed to laugh – "There's a ghost in the machine, isn't that silly?" But that mis-states the problem completely. It's true that Cartesian metaphysical dualism didn't outlast a century, but it wasn't because of problems with the ghost. The theory of mind, such as it was, was untouched. What Newton exorcised was the machine and not the ghost. In fact, what Newton showed, contrary to everyone's expectations (including his), is that you had a ghost all the way down, that is, even ordinary matter in the most elementary dynamics had ghostly properties. It isn't just the mind but everything is ghostly. Amazingly, Newton showed that contact mechanics just didn't work – it was true that you could influence things at a distance. Scientists of the day, including Newton, thought that this was absurd. Newton himself called universal gravitation an "occult" property. He said that anyone with any brains can see that it is inconceivable that something can influence something that it is not in direct contact with, and yet we seem to have to assume that. The major scientists of the day, like Huygens, threw this out for reasons that it was idiotic. Newton himself was torn by it, but he realized that it just has to be right, because it is a spectacular breakthrough.

On the other hand, it made no sense because it was inconsistent with mechanical philosophy. What happened was that Newton didn't get rid of the ghost in the machine but he just showed that all properties of matter are ghostly. It's all a ghost. It's all unintelligible, all the way down to terrestrial motion and planetary motion, which remains a big problem. It led to a totally new way of looking at science. I. Bernard Cohen (major Newton scholar) points out that "by entering into this paradoxical world, Newton set forth a new view of science, in which the goal is to find the best theoretical explanation, irrespective of any intuitive notion we may have of ultimate explanation." The point is that there is a ghost in the machine, and that's just the way machines are – they're ghostly – and we can't do anything about it. The mechanical philosophy appears to be wrong, although it is self-evidently true. From this point on, "we must be satisfied with universal gravitation and that it exists, even if we cannot explain it in terms of the self-evident mechanical philosophy."

That's just where we are. Science means accepting universal gravitation if we have evidence for it, and if it provides us with an intelligible scientific theory, even if there's no way of accounting for it in terms of what's self-evident, namely, the mechanical philosophy, which you just have to admit is wrong. From this point on, people's intuitions about what must be true become irrelevant. So, if the way to deal with universal gravitation is through curved spacetime, fine. If the way to deal with the universe is through weird quantum mechanical effects, fine. Too bad for our

intuitions. If the world really is made up of infinite one-dimensional strings in ten-dimensional space, okay, we're stuck with that. Time has a beginning? Fine, time has a beginning. Whatever lunatic idea people come up with tomorrow, it has to be evaluated on some merits – does it yield insight and understanding? Does it help us come to terms with the nature of the world? If it does, then it satisfies the conditions for rational inquiry, irrespective of our intuitions. The Newtonian revolution (even the earlier Galilean revolution), that's its essential content, I think. From then on, you are off on a totally new path. That's why it is the one real scientific revolution in human history – because it just set inquiry off on an entirely new path, and that's where we are now. We have no other criterion; common sense criteria are irrelevant.

Now, we can't get out of our skins. I mean, you go out in the evening and see the sun set. No matter what you know, you still see the sun set. And when you see the moon near the horizon, it is just bigger than when you see it up there. You can't *not* see the moon illusion. We see the world in terms of the way we are, and we can't help it. Just like we intuitively feel that the mechanical philosophy must be true, because how could it not be? But we have come to recognize that the way we see the world is just another fact to be explained about the world. If we see the sun set, if we see the moon illusion, if we believe in the mechanical philosophy no matter how much we try not to, then that's just a fact about the world – it's a fact about a very special part of the world, namely, the human mind-brain and the way in which it acts, conceptualizes, constructs, etc. And we would like to come to understand that, and in order to do it we have to make a kind of intellectual wrench – we have to take ourselves out of our skins and look back at ourselves, reflexively, as part of the world. That's hard to do, but we know that that is what it means to study humans naturalistically, and that's the move one must make. Therefore, the way we look at things in common sense no longer provides a criterion for intelligible explanation but rather is just a phenomenon to be explained.

Now, there is a real problem here which we can see right off. Whatever the right working notion of intelligible explanation is, its again something that's inside our skins, and we're stuck with that. That's something like a real paradox. Whatever our capacities to carry out rational inquiry may be in some corner of our brain (let's just call it 'the science forming faculty' to dignify ignorance with a title), it uses its own resources and its own criteria, which may be as misleading as the moon illusion. Maybe it's always systematically leading us away from the nature of the world. If we're creatures of the world, that would not be in the least surprising. But there we can't do anything about it because we're stuck with it, that's as deep as we can go. If we are systematically being misled about the world because of the way our science forming faculty happens to be constructed, we can't do much about it. But we can

overcome the moon illusion. It's hard but we can overcome it. We can't stop seeing it, we can't stop seeing the sun set, we can't stop believing in the mechanical philosophy (Newton's intuitions are our intuitions), but that part you can throw out. The part that's guiding our search for intelligible theory, we can't throw that out. Here we are approaching something which is kind of like a classical paradox maybe, but we have to recognize the naturalistic point of view that we are just a particular organism trying to understand the world, and we have to do it our way because there is no other possibility.

We have now abandoned metaphysical dualism, but abandoned it in a very special way. It's not that there's a ghost in the machine, it's that the machine has ghostly properties all the way. Maybe it has even more ghostly properties above the neck. If so, that's just another fact. However, it is already unintelligible to common sense, down to elementary dynamics. That's Newton's basic discovery. That's why contemporary discussion ridiculing ghost in the machine is completely off-track. It's just missing the point of what happened. There was no criticism made of the Cartesian theory of mind. You could argue that that was because it wasn't substantive enough. Maybe. But it any event, it survived all of this intact, such as it was. What did not survive is the theory of the machine. Contact mechanics was thrown out, and the common sense of the next scientific generation is that the mechanical philosophy is wrong.

It took a long time for this to settle in. For a couple hundred years, physicists were still trying to find mechanical explanations, and it really wasn't until the 20th century that it was given up. You get as late as to people like Henri Poincaré, who was arguing that the molecular theory of gases we adopt only as a computing mechanism, and the only reason we adopt it is because we're familiar with the game of billiards. Efforts to try to explain things in terms of the ether and stuff like that were an attempt to carry out the mechanical philosophy. It was really in the 20th century before Newton's insight was basically incorporated into the sciences, and it may be even harder to do it in the emerging sciences. Anyhow, that's the way we must look at it in retrospect.

I'm being kind of anachronistic when I say that the 17th century exorcised the mechanical philosophy – it did it in principle, it didn't yet do it in fact. It left a big residue of unhappiness and confusion that took centuries to sort out. But looking back, it seems fair to say, abstractly speaking, that Newton exorcised the machine. He got rid of the mechanical philosophy. It's anachronistic to say it but it is an accurate historical reconstruction, and I'll continue with this historical reconstruction.

One consequence of eliminating the machine is that we have gotten rid of any notion of body, physical, or material. There no longer is any notion of the material world. Remember, Descartes could be a dualist because he had a notion of body. Not a very clear notion, but at least a very general notion – 'body' was defined by the mechanical philosophy. Newton showed that that's not the way body works, therefore we have no concept of body – it is just the world, whatever it is. If the world has mental properties, then that's part of the world. But there is no notion of body, unless somebody comes up with a new one, and nobody has ever done that. Unless somebody comes up with a new notion of physical, talk about 'physicalism', 'materialism', 'eliminative materialism', 'the mind-body problem', etc., just seems to be meaningless. If this is correct, then something very curious has happened, because people are talking about it all the time. There are all kinds of stuff written about the mind-body problem right up to the present. There are theories in the cognitive sciences, say, of the Churchlands, called 'eliminative materialism', which says that we shouldn't study the mind but we should study the material world. There are lots of questions about 'physical reductionism', i.e., can we reduce things to physical terms? This is just universal in philosophy, and I don't understand it. Maybe somebody can explain it to me. It seems that once we have lost the notion of physical, none of these questions even mean anything. You can't have the problem of reducing things to the material if there's no notion of material, and there isn't any. Somebody could say that 'material' is what they teach in the physics department, but even in the physics department that's not true, because they expect in two years from now they'll teach you something else, at least if the subject is still alive. So, that can't be true, and there's no other notion – all we have is the world with whatever properties it has, and there's no notion of the material world, and hence all problems of reduction disappear. The problems of elimination of the study of mind in favour of neurophysiology don't mean anything. They're just purely irrational in saying, 'Instead of studying this part of the world, let's study that part of the world.' It doesn't make any more sense than saying that since solids are hard to study, let's study liquids. It's of no more interest than that, unless somebody can come up with a concept of the material to replace the Cartesian concept (which has been thrown out the window). As far as I know, nobody has even addressed that problem, let alone offered an answer to it. And if so, virtually all the discussion in the field is not just wrong but literally meaningless – it's talking about a problem that can't be stated. That's the strongest version of the position I'm trying to take. Let me put in the most outrageous form possible – the whole theory is not wrong but meaningless. It's talking about something that it can't characterize, namely, something that presupposes the notion of physical that has been abandoned in the 17th century.

With metaphysical dualism now unstateable, notions like eliminative materialism lose sense. The natural conclusion to draw from Newton's demolition of the theory of matter is that human thought and action (like the creative aspect of language use) are just properties of organized matter and nothing else. In fact, this conclusion was sporadically drawn not too long after Newton. It was first drawn by La Mettrie about half a century later, and it was considered totally outrageous. In fact, his work wasn't even revived until this century. And La Mettrie was driven out of France, and out of Holland, but he finally survived thanks to the protection of Frederick the Great for some reason.

About a generation later, in more tolerant England, the same idea was developed by the eminent chemist Joseph Priestley, a major 18th century scientist, who argued that thought in humans is a property of the brain – a necessary result of certain organization. It's like electricity, magnetism, and the powers of attraction and repulsion. Nobody really understood those things then, but whatever electricity, magnetism, and the powers of attraction and repulsion are (these ghostly properties that matter has), thought is just another one of them, and we have to investigate it. Another 18th century version of it is that – the brain secretes thought the way the liver secretes bile. That's kind of the image.

If you look at people like John Searle today, he is supposed to be saying something more or less similar to that effect (if I understand him correctly). Well, that looks like the right move, mainly because there's no other choice – there doesn't seem to be any other move we could make. If all there *is* is the world, which has ghostly properties all the way down to elementary dynamics, then there's nothing to say except that the properties of attraction, repulsion, electricity, magnetism, quantum effects, thought, etc., are just some properties of the world, some property of the way matter is organized, where matter now has no meaning – matter is just some property of how whatever there is that's organized. And a naturalistic approach to studying humans ought to proceed that way.

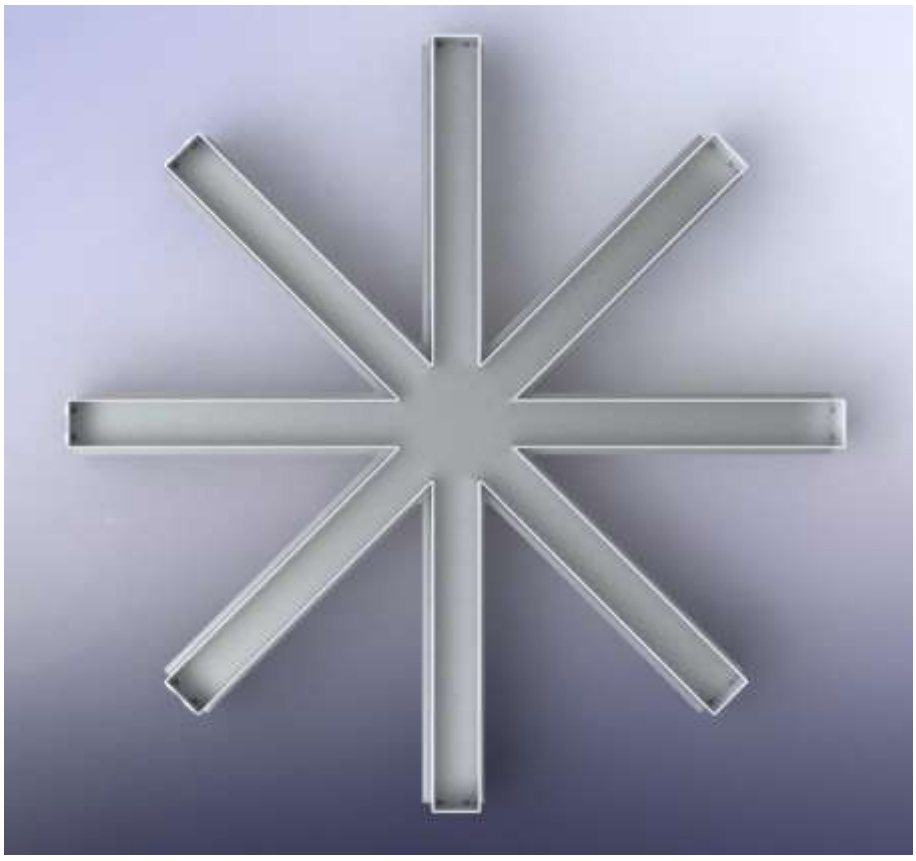
That leads us to the next question – how can organized matter have these properties? On that question, progress has been essentially zero. We have nothing to say about how organized matter can have such properties as the creative aspect of language use. In that respect, we're exactly as much in the dark as the Cartesians were. It's not that matter and mind are different kinds of things, because there doesn't seem to be any notion of matter, but if we think of those terms as descriptive conveniences (describing certain aspects of the world, viz., humans above the neck vs. everything else), then it seems that matter and mind pose different kinds of problems to human intelligence. That appears to be the case. Maybe that'll be shown

to be wrong, like, there was a time when electricity and magnetism (or, for that matter, universal gravitation) also posed total mysteries, but maybe we at least have learned enough. But for now, it seems reasonable to suppose that matter and mind, though not different kinds of things, do pose different kind of problems to human intelligence.

If that turns out to be true, then it would not be in the least surprising to a naturalist. Remember, a naturalist's perspective assumes that humans are just part of the world and are not angels. If humans are a thing in the world, they're going to have certain cognitive capacities, like rats, for example.

Rats can do certain things but can't do other things. It's very lucky for a rat that it can't do lots of things, because if it wasn't incapable of doing lots and lots of things, it wouldn't be capable of doing anything. There's a logical relation between scope and limits. If you don't lack the capacity to do many things, you can't have the capacity to do anything, for a very simple reason – the capacity to do something requires articulated structure, and if you have that then it's going to rule out all sorts of other things. The point is obvious in the case of physical growth – an embryo has very rich genetic instructions allowing it to become a chicken, but that very structure will prevent it from becoming a monkey. So, lucky for a chicken that it lacks the capacity to become a monkey, because if it didn't lack that capacity, it couldn't have become a chicken. And assuming the world doesn't change somehow when you move to cognitive function, the same thing is going to be true there.

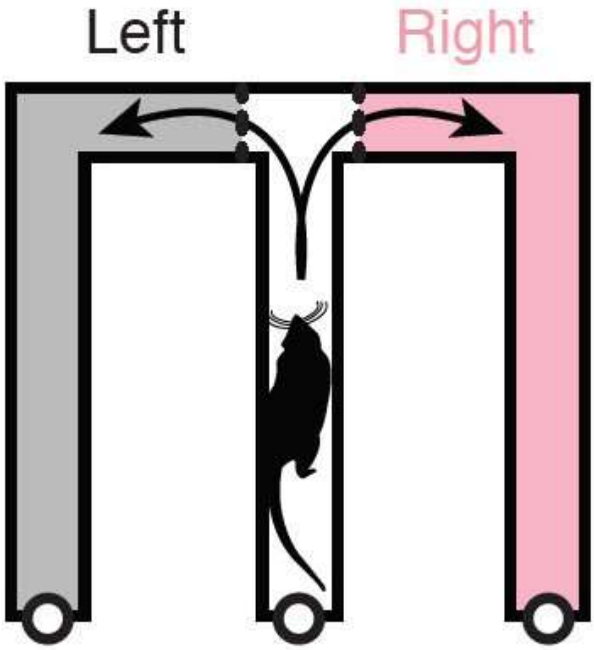
So, rats happen to be very good at solving what are called radial mazes.



Credit: animalab.eu

A thing like this where you stick a rat in the middle and keep some food in one of the spokes, and the trick is to learn how go down each path exactly once, like, remember you didn't move down that path before. And in order to make it as hard as possible, the experimenter will rotate the thing so the rat can't have any olfactory cues. It turns out that rats can solve this very fast, probably faster than humans. So, they're very good at radial mazes, but they're horrible at other mazes.

For example, take a right-right-left-left maze.



Credit: Kay et al./Cell 2020

We think rats can't do these at all. If they can, it will be extremely hard, whereas humans can do it easily. If you were to set up a prime number maze, where the rat is to turn right at every prime number of branch-points, obviously the rat could never do it in a million years, and the reason is that it just doesn't have that concept in its head. A rat can do many things we can't do, like build a nest, find its way home, solve radial mazes, and all sorts of other things, because it has special capacities. But it can't do things like a right-right-left-left maze, which humans can do. And it's lucky for the rat that it can't do these things, because that means it has the structure to do all these other things.

To introduce some terminology, we might distinguish for a rat *problems* and *mysteries*. By problems for a rat, I mean intellectual challenges that can in principle be resolved within the rat's cognitive space. It may take a long time, but it's got the concepts for it. By mysteries for a rat, I mean things that are outside its cognitive space altogether, like a prime number maze or a right-right-left-left maze. The analogy in embryology would be that becoming a chicken is a problem for a chicken embryo, becoming a monkey is a mystery. You can't change the nutritional environment of the cell in order to make a chicken into a monkey. Actually, nobody knows that, but it's just taken for granted – rational people looking at the world just assume that, although they have no real knowledge to explain it, because it's so obvious that nobody even talks about it. I'll come back to that fact, because it is

interesting that in the study of cognitive development people aren't rational the way they are in embryology. This irrationalism is part of the epistemological dualism that I'll come back to.

In areas where we can all be rational without too much trouble, like chicken embryos and rats, the distinction between problems and mysteries is quite clear. Notice that it doesn't have to be a sharp distinction. There could be grey areas, but as a first approximation it makes sense to distinguish these two categories, and it may even be very sharp. If humans are part of the natural world, then the same is true for humans too – there will be problems and mysteries for humans. There will be intellectual challenges that will be within our cognitive reach in principle, and there will be some that are not. If that is true, then there are certain things we can't ever understand, and we should be very happy about it because that's a consequence of the fact that we can understand anything at all (for reasons of logic).

In theory, we can find out something about our problems and mysteries. Abstracting away from our skins and looking back at ourselves, we could carry out an investigation that might set the boundaries of those things. For example, we might discover that there are certain kinds of intellectual constructions that humans are capable of setting up. If you look at science, there are certain ideas that keep cropping up – we could study input-output systems, we could study deterministic systems, we could study probabilistic systems, etc., and if something can be put into those frameworks, then we can deal with it. On the other hand, it may very well be that the creative aspect of language use just doesn't have those properties. It's an aspect of the world that just lacks those properties, exactly as the Cartesians thought, in which case it will just be a mystery for us (like free will and stuff). Same might be true of the problem of consciousness. In fact, all the old chestnuts, all the questions everybody is worrying about for thousands of years, we have never made any progress at all. People don't even have bad ideas about it. It just looks like you're pumping into a blank wall. You know as well as anything that you're aware of some things and not of others. You know the Cartesians were right in saying that whereas a machine is compelled, humans are only incited and inclined, and could decide to resist those inclinations. You just know that as much as you know anything, and it still just looks like a total mystery. There isn't any idea, however bad, as to how that might happen. There are ideas about quantum mechanics and so on, but they don't even reach the level of bad ideas, because they always include random elements, and we know that the creative aspect of language use is not random. It's no more random than it is determined. It is somehow appropriate but not random. In fact, it's just uncaused (not uncaused in the sense of 'random' but uncaused in the sense of 'undetermined but appropriate'). There's no problem recognizing the phenomenon,

but we haven't the foggiest idea what it is. It could be that we're facing something like a right-right-left-left maze for a rat. It could turn out to be true that the domain of the mental for humans is just a mystery, and hence not metaphysically different (as there are no metaphysical differences) but epistemologically different – it's part of our mystery space and not problem space.

There's a book coming out by an Oxford philosopher Colin McGinn who argues that philosophy is just the study of mysteries in this sense. That's why philosophy is hard, because it's the study of mysteries. You can never make any progress. Actually, there's a traditional view of philosophy which would make this almost tautological. One traditional view of philosophy holds philosophy to be the 'mother of the sciences', meaning things are philosophy if you don't understand them, and as soon as you begin to understand them, they become science, and philosophers don't worry about it anymore. In the contemporary period, people like John Austin strongly advocated this view and said that what he was working in was pre-science (the theory of speech acts), and whenever they understood it enough, it would go off to become part of the sciences, and then they wouldn't talk about it in the Oxford common rooms anymore because they'd really work on it. If that's what philosophy is, then by definition it is the study of mysteries, so it's problems really are kind of hard in some special sense.

Notice that what's a mystery for a rat may not be a mystery for us, and conversely. The notions of problem and mystery are organism-dependent. There's not absolute sense in which something is a mystery. There could be a Martian for whom what's a mystery for us could be trivial. Maybe he's watching us all the time and wondering why we're always making the same dumb mistake when we study mental phenomena, just as we watch a rat and wonder why it always makes the same dumb mistake when it's running a right-right-left-left maze. And we'd think the mysteries to the Martian might be trivial to us, there's no commensurability necessary in this space. Something like that must be true from a naturalistic point of view, if humans are a part of the natural world and not angels. We don't know what the boundary is, we just know it's there.

Coming back to the question of the natural sciences, it would seem that the natural sciences are an accidental convergence, a chance convergence between some properties of the world and properties of our cognitive space. There's no reason why it cannot be like that. There's a long discussion in the philosophy of science trying to argue that it had to be like that, which goes back first to Charles Sanders Peirce about a century ago. Scientists have always been asking the question of how science is so successful. It's kind of a miracle. Not foolish scientists like Einstein asked this

question. How come the human mind is capable of understanding the nature of reality? Of course, that presupposes that humans *are* capable of understanding the nature of reality, and the evidence for that is very slight. In fact, what seems to be the case is that through the general obscurity, there are a few little points of light that have broken through and that's what we call science. But most of the questions (say) the Greeks asked are just as obscure today as they were then. Only a small number of them have been answered.

Firstly, the fact to be explained doesn't seem to be a fact. The fact that humans have this amazing capacity to understand the nature of the world just doesn't seem to be true. Most of what's going on in the world we haven't a glimmer of understanding about. So, since the fact isn't a fact, we don't care what the explanations for it are. But the explanations that have been offered, from Pierce up to people like Stephen Hawking in his recent book (*A Brief History of Time*), are always the same mistaken explanation, namely, that the reason for this non-fact is evolution – evolution selected us to be able to solve problems and therefore we're able to solve problems. That doesn't make any sense at all. There was nothing in the history of human species that gave a reproductive advantage to the ability to solve problems in (say) number theory or quantum theory. In fact, human evolution took place in hunter-gatherer societies, and nothing has happened since then. There's no reproductive advantage, even in modern societies and certainly in hunter-gatherer societies, to solve problems in the advanced sciences. And that's all evolution is about – reproductive advantage, nothing else.

Serious evolutionary biologists, who have been trying to ridicule these notions for years, have even argued only half-jokingly that there's a selective advantage for stupidity. Richard Lewontin, one of the main modern evolutionary biologists, trying to debunk all kinds of stories about cognitive development and so on, suggested that if we really look back at hunter-gatherer societies and consider the people who for some genetic reason were more adventurous or more imaginative, chances are that they will get killed, although they will be good for the tribe. That is, when all the dumb, unimaginative people are sitting around the fire, this imaginative person is going to be out trying to figure out a way to hunt the sabretooth tiger. Maybe this person will succeed now and then, in which case it is good for the people sitting around the fire, but over time it tends to be bad for him, because he often fails and the tiger gets him. Therefore, there is a reproductive advantage for being passive, stupid, unimaginative, etc., and therefore, Lewontin argues, there must be an evolutionary tendency in that direction. This is only half-serious (although not entirely unserious, incidentally), and merely an effort to debunk the idea that you can learn anything about the evolution of cognition. Lewontin's major point is that

there's nothing possible within any scope we can dream of that anybody will ever be able to say about the evolution of the language faculty or the evolution of cognition, and certainly not about the evolution of our science forming faculty. It's just hopeless. It's just some physical configuration that took place for whatever reason. And I think Lewontin is correct. It doesn't mean that there isn't a lot of work on it, but the work seems vacuous to me, for those kinds of reasons.

In any event, it's reasonable to suggest that science, as we know it, is just some ray of light that accidentally happened to break through because of some chance convergence between our problem space (which is part of our nature) and the world (which is whatever it is). And already by the 17th century big paradoxes were arising, because our problem space is wedded to the mechanical philosophy, and we had to give it up and move to some other parts of our mind, which have whatever properties they have, and which may be systematic in misleading us, and may be leading us away from inquiry into the mental.

Suppose certain aspects of the mental like the creative aspect of language use (the litmus test for the mental for the Cartesians) really turn out to be mysteries for humans, then we're just stuck with it the way we're stuck with the moon illusion. And maybe it is a mystery not only for our perceptual space but even for our cognitive space, in which case we will never have any understanding of it, just as we have no understanding of most problems. That doesn't mean we can't deal with these things in our normal lives. In our normal lives we're always dealing with things more or less successfully, although we haven't any conception of it. We do it by what's called our 'intuition', which is just a name for something we can do but without any understanding of how we're doing it. Virtually everything that goes on is by intuition. Fortunately, we've got all these capacities (whatever they mean), but to gain a scientific understanding of them is kind of unlikely. You might even argue that that's a good thing – it's nice that we don't understand ourselves too well, because it could be awful if we did understand ourselves too well.

Anyhow, it just looks plausible to assume that most of the humans-above-the-neck phenomena we'll just never understand. We'll be able to do all kinds of things but we won't be able to understand it, or at least it looks like that, and certain aspects of the mental appear to look like that. They just appear to be in the mystery space. Let me stress again that this could be a research topic, and in fact it could come out of cognitive science, that humans are capable of constructing certain kinds of conceptual structures (like determinism and randomness), and those are going to deal with certain types of phenomena, and if you can show that certain arrays of phenomena don't have those properties, too bad, then we're in the mystery space.

That's a conceivable empirical discovery about human beings and there's nothing paradoxical about it – we could discover what is a mystery for us. We can't solve it but we can find out where it is. A large part of the so-called mind-body problem could ultimately fall into that area. Again, this is not metaphysical dualism because we can't state the mind-body problem.

With metaphysical dualism now unstateable, is there another kind of dualism? Remember, metaphysical dualism falls within naturalism. It's just a form that naturalistic inquiry takes. Naturalistic inquiry reaches metaphysical dualism if it takes the road the Cartesians followed, and then the dualism explodes because the world disappears, the machine disappears, and now we're left without any form of metaphysical dualism, since there isn't any matter anymore.

Is there any other kind of dualism that we can construct? The only other kind is the irrational dualism of the epistemological kind which says that we are just not allowed to study humans above the neck naturalistically. As distinct from metaphysical dualism, this is a completely irrational position. As far as I can see, it has no saving graces at all. What I want to show is that however irrational it may be, it is pervasive, and in fact it dominates (maybe universally dominates) the more reflective, considered aspects of the cognitive sciences – by that I mean the contemporary philosophy of mind, philosophy of language, AI, debates about the limits of machines, questions like 'Can software think?', etc. What I want to try to argue is that if we look at it closely at each point, all of this is, ultimately, a departure from the naturalistic approach to humans, meaning it is a form of epistemological, hence irrational dualism. There might be an argument for treating humans non-naturalistically, but then the burden of proof is on anyone who suggests that. If you can't meet the burden, then we can forget about it.

That's what I'd like to argue, and to the extent that it's plausible, the last question (which I'll talk about at the end) is that why would it be true that there would be a natural explanation, namely, that intuitively we're just dualists, and we can't help being dualists any more than we can help seeing the sun set. Probably, our approach to humans is just irremediably dualistic – we see humans as having ghostly minds in a physical body, and even if the notion of the physical body disappears, we can't help seeing people that way, just as once contact mechanics and the Ptolemaic universe disappeared, we can't help seeing the sun set.

If this is true as an aspect of human cognitive character – if metaphysical dualism is on a par with the setting of the sun – then we're forced to irrational

dualism, with metaphysical dualism gone. But that's a path we should resist. That's the diagnosis I'd like to give at the end, but first we have to traverse the path.

Lecture #2

I SUGGESTED THIS morning that in a post-Newtonian world, we are left with no notion of matter, body, and physical, hence no notion of physicalism, physical reductionism, eliminative materialism in the sense of many cognitive scientists, and, of course, no notion of metaphysical dualism. These notions simply seem to have no definitions. They don't have any meaning because the notion of physical has disappeared; because Newton demonstrated that the physical has ghostly qualities all the way down to the simplest phenomenon.

Either this is true or this is not true. If it is not true, then I'd like to see a reason for it. If it is true (as I think it is), then it's taken much less seriously than it should be, because if it is true then it's extremely hard to translate an enormous amount of discussion in the philosophy of mind literature into something that makes any sense. The discussions just don't seem to mean anything, because there doesn't seem to be any topic. There would be a topic only if the notion of physical were characterized in some manner, and that no one has tried to do, and they haven't tried to do it mainly because they know it can't be done. The physical is just whatever we come to understand more or less, or the physical is whatever that is actually *there* (if you want to take a realist position), and there's no question of reducing anything to *that* because everything is just a part of *that*.

With metaphysical dualism now unstateable (and recall that metaphysical dualism could be (and in the Cartesian version was) a naturalistic position), is there any alternative to naturalism? Is there any kind of dualism that one can put forth? There is only one kind – a methodological/epistemological dualism – that says that we're not allowed to study humans the way we study everything else, or we are, at least, not allowed to study humans above the neck (the mental) the way we study anything else. As distinct from metaphysical dualism, this appears to be a totally irrational position. It doesn't have any saving graces. Unless somebody gives a reason for it, it seems that it should just be dismissed. Nevertheless, as I mentioned, it's a very pervasive position. It not only is pervasive but all-inclusive. I'm referring now to the considered parts of the field (not the actual research but the thinking about it), which fall under irrational, methodological/epistemological dualism in this sense. I'll give some examples of that in a moment. But, for concreteness, let me just spell out what I think a naturalistic program would lead to. There's only one area of

the mental where there's real progress, namely, language, and let me sketch out what a naturalistic program would lead to in the case of language.

In the case of language, what one finds is that what's involved in human language is the human brain and not (say) the human foot. If you cut a person's foot – no trouble thinking. If you cut out the brain – trouble thinking. To start with, language seems to be localized in the brain. Furthermore, it seems to be in certain parts of the brain. That doesn't necessarily mean an area you can cut out, although even that's partially true, but certain subsystems of the brain seem to be specifically dedicated to language – so let's just give them a name and call it 'the language faculty.' When we look more closely at the language faculty, we find two kinds of systems – the cognitive system and the performance systems (a class, actually). But the world doesn't have to be like this, and if this is true, then it's an empirical statement about human language and the architecture of the brain.

The cognitive system is a computational-representational system that stores information. It, in fact, stores infinite information that we all have – about the properties of the expressions of our language, about their sound properties, meaning properties, structural properties, etc – which is to say that our cognitive system is a finite object that stores an infinite amount of information. That notion was a big problem throughout modern history, when people were thinking about this topic. It was recognized long ago that we have unbounded knowledge but finite minds, and that was an insoluble problem until the 20th century when it finally became very clearly understood what that means. What that means is there's a certain kind of property of the brain, namely, the property of being what's called the generative (or a recursive) procedure, which characterizes an infinite amount of information in a finite way. That happens to be a particularly well-understood property now. Hence, saying that the brain has this property is not mysterious. It could be false, but there's nothing mysterious about it. Surely, the brain has properties, and this particular property – the property of being a generative procedure – is particularly well-understood. There's a whole field of mathematics where it's very well understood. With that formal insight, it becomes possible to address the traditional concerns about how you could have infinite, unbounded knowledge with only a finite thing. And, in many ways, modern linguistics comes out of that confluence – the confluence of a formal insight and a traditional problem.

Now, unfortunately, it didn't work like this. In the rational world it would've worked. The reason it didn't work was that the traditional concerns had all been forgotten by the time modern linguistics came along. We were in an externalist-behaviourist world in which the traditional concerns disappeared. So, you had to

reinvent traditional concerns as well because of the remarkable parochialism of all the professional disciplines – including linguistics, philosophy, and psychology – none of which were aware of the traditional concerns.

I should say that these traditional concerns went not very far back. As recently as the 1920's, Otto Jespersen, a well-known Danish linguist, coming now at the tail-end of a century-old tradition, wrote that the goal of linguistics is to characterize what he called the notion of structure that people have in their minds that enables and guides them to form and understand arbitrary expressions, including what he called 'free expressions' – ones that they had never heard or produced before. That's the right idea, and it goes back a couple hundred years, but when I was a student in the 1940's, nobody would read Jespersen. In fact, I learned about this as a curiosity, looking up books in the library. And everything else in history disappeared as well. The schizophrenia of serious scholars around this period (the 1940's) was quite remarkable, in retrospect. There are lessons in this for the future, which is why I am going off in these anecdotes.

The leading American linguist of that period was Leonard Bloomfield, who happened to be a Sanskrit scholar and a scholar of Indo-Germanic studies, which was one half of his brain. In the other half of his brain, he was trying to be a hard-headed scientist – reading stuff from the Vienna circle on logical positivism, contributing to its meetings, etc. He was laying the foundations for modern hard-headed structuralism – he was a committed behaviourist. He was considered, at the time, a tough-minded scientist. If you look back at his work, there are two kinds of work, which he kept totally separate. How he managed this internally, I have no idea.

In the Germanic-Indic-Sanskrit side of his brain, he was writing generative grammars. He wrote a detailed grammar of an Algonquin language called Menomini in the 1930's, with rules, rule-ordering, etc. Though it didn't go into syntax, because no one knew how to bridge that gap between infinite information and finite rules, he did deal with parts of language that people knew how to finitely characterize (phonology, morphology, etc), and he dealt with them by ordered rules and all sorts of things that are by now familiar. The reason he could do all that is because he knew traditional Indian linguistics from 2500 years earlier, when that was exactly the way it was done. If you read, say, Panini (5th century B.C.), it is a generative grammar in the modern sense.

On the other hand, as a theoretician and a hard-headed scientist, Bloomfield was writing derisive critiques of crazy mentalists talking about ordered rules, when

everybody knows that all there exists is behaviour and the structure of discourse and so on. It's interesting to note that his *Menomini Morphophonemics* he published in a Czech linguistics journal, something nobody in the U.S. would ever read. I was a student about ten years later and my professors were students and associates of Bloomfield's, and one of them, in particular, had the same mixture of interests – he was an Indo-Germanic scholar who knew Sanskrit and was also a modern linguist. At that time, I also wrote a generative grammar as an undergraduate thesis, but nobody ever pointed out to me that Bloomfield had done it ten years ago. They had to have known it. I found out about this years later when all this stuff was rediscovered. The point is that his generative work was so deeply hidden that even people who had to be aware of it never recommended it to an undergrad, because it was considered so outrageous. Then why did Bloomfield do it? Because in the sensible half of his brain he knew there is no other way to do it. But what was sensible was totally repressed.

This is a warning about excessive parochialism and about taking seriously things that ought to be questioned, which in an area like this means almost about anything.

Coming back to our picture – we have a cognitive system, which is some generative, recursive procedure that stores an infinite amount of information, particularly information about sound-meaning relations. As to the system's nature, we have a huge amount of fairly reliable data from a lot of different languages, and some theories which have gone somewhere toward explaining a fair amount of evidence. Though not a lot is known about the cognitive system of the language faculty, there has been substantial progress in understanding, and everyone sort of knows how to proceed, and there's plenty of work to do. So, it's a normal science problem in that sense.

With regard to the performance systems, much less is known. It's generally assumed that the performance systems are fixed and invariant – that they don't change through childhood, that you don't learn them. It's usually assumed that you don't learn, for e.g., parsing theory – you just have it. The same holds for every other performance system, and the reason for that assumption is not that anybody knows anything about it, but precisely because nobody knows anything about it. Since you don't know anything about it, you might as well make the simplest assumption, which is that it is invariant, until you have some other evidence. The usually tacit assumption that performance systems are fixed and unlearned is basically due to ignorance. It could be true or not be true, but until more is learned about it, it will be unanswered.

With regard to the cognitive system, which is better understood, it is known that it is not fixed, or at least not entirely fixed. So Catalan is not English, and I'll find that out if I walk outside, although that could very well be a misleading impression. For a rational Martian scientist looking at the earth, he would probably say that Catalan is just English with some trivial changes that aren't worth looking at. It is reasonable to suppose in advance that that's what will ultimately be discovered. The reason it is reasonable to suppose that is that if that isn't true, then it is extremely hard to explain the fact that anybody knows any Catalan or English, that is, if you compare what is known with the database for it, the gap is so extraordinary that, short of miracles, you can only conclude that it all came from the inside. But if it all came from the inside, and since it is obvious that we're not genetically designed to speak English or Catalan, then it just has to be the same language with minor variations. And one of the intellectual challenges of the field is to demonstrate that what you know in advance has to be true – that there really is only one language, and that the differences between them must be very peripheral and must be located in those parts of the system for which there is direct data. So, about phonetics there is a fair amount of data – you hear it – so that can vary a little bit. But for most of the computational part of the system, you have no evidence about, and therefore it almost has to be unique.

Current theories do propose that there really is only one computational system for language, and that the differences are located completely outside the computational system. That's far from having been proven, but that ought to be true, and you can now postulate it without absurdity, in that you know enough so you can see how it could be true, and assuming that it's true helps explain a lot of things. But that's getting ahead of the game.

In any event, the cognitive system does undergo some changes that are probably limited changes. It starts with some kind of initial state, goes through a series of states, and ultimately stabilizes. Apparently, there is a 'critical period' for this. It is not certain, but there's mounting evidence for a critical period, meaning that it has to take place during some period of physical development –probably pre-puberty. It seems that the changes that take place later are much more peripheral to the system, but the major fixing of the system seems to be extremely early. In fact, the better the experimental techniques are, the earlier it gets. By now, some of the results are pretty astonishing.

For example, there's some evidence that a four-day-old infant can distinguish the native language of its mother from another language spoken by another person. This experiment happened to be done in France. So, if a four-day-old French infant

is listening to a bilingual woman (who knows French and Russian) not his mother, the infant will respond differently to her speaking French than to her speaking Russian. That means that within four days, or even before birth, some differences got fixed. Things like the basic intonational structure of languages seems to be at least partly fixed before six months of age, i.e., before the kid has even said one sound. The same seems to be true of things like finding the regions for which certain phonological units appear – like the regions where (say) a vowel might differ somewhat from Polish to Chinese. Those things also seem to be fixed pretty well by six months, and if we get better experimental techniques we may even go down further. One of the exciting things about the work in developmental psychology in the last couple of years is that with the improvement of experimental technique, you just keep pushing down the stage of innate knowledge earlier and earlier, and by now in most areas it's way before any overt behaviour. It's well-established by now that for 20-month-old children (whose behaviour is what are called two-word sentences), they can not only fully comprehend seven- or eight-word sentences, but even figure out the meanings of words from the structural properties of those sentences – all of it way in advance of any performance.

All of this shouldn't surprise anyone who takes a naturalistic point of view to humans. After all, we assume that automatically for every other aspect of human development. Nobody assumes that other aspects of development are taught. But for the study of cognitive processes, this is an important discovery, because it tends to overcome the irrational dualism that I've been talking about. In any event, the cognitive system does change to a certain degree, but it doesn't seem to change very much. It gets fixed at some stage – apparently, pre-puberty (maybe well earlier than that) – and within pretty narrow limits. The attained state (in fact, all the states) is a generative procedure of the kind that I mentioned – a computational-representational system. A generative procedure generates something – it forms things. In this case, the generative procedure (this is what's sometimes called 'grammar', but that's a bad term so I'll keep away from it) generates a certain set of objects called 'linguistic expressions.'

GP → {linguistic expressions}

Each one of these objects is in itself a collection of properties. For example, one of them is the word 'the' in some sentence or other. That sentence or other is some collection of properties, namely, a collection of phonetic, semantic and other structural properties. Hence, each linguistic expression is a collection of a variety of properties. For example, the linguistic expression 'table' in my language has certain sounds and certain instructions as to how to use it to refer, etc. The generative

procedure assigns those sets of properties to every expression. In other words, what's generated is a set of expressions, technically called 'structural descriptions.' We can pull out some piece of these things. For example, if we look at only the phonetic part of the linguistic expression, we have something that we could give a name to, like a 'signal'. Similarly, if we just pull out semantic and structural parts, we would have something that we might call a 'concept' or something. But the linguistic expression itself is just a collection of all these things.

The signal could be phonetic, but that's a little misleading because we now know it could be other things – like sign, for example. Sign seem to be learned pretty much the same way a spoken language is, and it seems to be localized in the same parts of the brain, which is quite surprising because the modality is completely different – it's coming in through a visual rather than an auditory modality – but it seems to end up being processed in the same part of the brain –actually, the left side (the side that's not used for visual processing) – which is a particularly remarkable fact about sign language. And there's a lot of interesting discoveries about that which tell you a lot about the structure of the language faculty of the brain and how its parts are parcelled out for different kinds of activities, etc.

In any event, the cognitive part seems to be sort of modality-independent. It doesn't seem to matter whether the input to it is coming in through the eye or the ear, as long as it is of the right type. That means that the phonetic part of the linguistic expression must be understood somewhat more abstractly, as whatever part that gives instructions to the production and perceptual systems. In normal cases, that part is articulatory. But it's now known that that's not the only case, and how many others there are, nobody knows. You don't do experiments in these things; you just look at what's around. That Martian I was talking about might do experiments on us and could learn a lot, but for ethical reasons we don't do it.

The linguistic expression then is a collection of those things, and it has manifestations – a signal is a manifestation. In a much more general sense, you could say that speech acts are also manifestations of linguistic expressions. If you think of the linguistic expression as a set of instructions to performance systems (for example, for the articulatory system), you'll end up with what the signal will be. If you look at a broader set of instructions to other performance systems, you'll get an act that a person performs (like the act of referring to a table through the word 'table', which is something I could do), and that act is another manifestation of the linguistic expression, where the instructions are being used by other performance systems. That's the basic picture.

Now, here we must begin by being a bit careful. It's common in just ordinary discourses (in the unreflective talk about language) to say things like "This word has changed its meaning in the last 100 years", "It has changed with sound in the last 100 years", "It used to be pronounced this way, now it is pronounced that way", or "It used to mean this, now it means that." Strictly speaking, all that talk is meaningless. The word is just a collection of its properties, and if the properties change, then it's a different word. It is as if you pick some mountain (say, Mount Canigou, if that's around here somewhere) and you asked if it is the same mountain after an avalanche. That's a matter of decision and not fact – it's the same mountain if you want to call it the same mountain. Is it the same mountain if one grain of dust is removed? Again, that's a matter of decision. Within a certain range of human interests, yes, it is the same mountain. Within another range, no. Similarly, if I say that there's no absolute answer to this question, that means that there's no absolute meaning to the notion 'the word changed sound' or 'the word changed meaning'.

There isn't much confusion about this in connection with words changing sound, but the question of words changing meaning is a big topic. An awful lot of philosophy is concerned with how words get meanings fixed. That's no more significant than the question of how words get their sound fixed. The latter is not considered a big problem because we are less confused with sound than with meaning for some reason, but it's basically the same question, viz., how the collection of features of a particular word gets combined, and that question is neutral between sound and meaning. To say that words change meaning or to say that there's a certain way to fix their meaning is to really not ask a very serious question.

Similarly, when you ask whether a word means the same thing for me as it does for you, or whether it sounds the same for me as it sounds to you, that doesn't, strictly speaking, mean anything, because it's not the same word for you and for me. It's like asking whether you look like me. Again, there's no absolute answer to that. Relative to a certain set of interests, you might say that Luigi and I look alike (compared with somebody from Africa, say), but relative to another set of interests we don't look alike. There's no true answer to the question of whether we look alike. The same is the case when you ask whether two persons have the same word, the same expression, the same language, or whatever. Although there's nothing very profound about this observation, it's worth keeping in mind because it's constantly forgotten. I'll come back to cases where it's forgotten.

A lot of the contemporary theory of reference is called the 'externalist theory of reference', the idea being that the reference of words is determined by things in the world and communities. That's the reigning theory of reference today. I think an

awful lot of that is based on this confusion, and it disappears as soon as the confusion is eliminated.

Before going to the applications, let me just repeat what should be obvious – that a word, a phrase, a sentence, a text, etc., is a collection of properties, some of which are instructions to the articulatory system, some to the referring system, some to some other thing, etc. That collection of properties is the linguistic expression, and there's no substantive meaning to the question of whether one person has the same linguistic expressions as another. It is like the question of whether they have the same shape – yes or no, depending on what you're interested in.

Suppose that the cognitive system does pass through a sequence of states and hits a sort of stable state. We will assume, mainly due to ignorance, that the performance systems don't do this. Suppose the cognitive system reaches a state L, where it sort of stabilizes in some early stage of life (maybe necessarily in the early stage of life), and after that it undergoes only peripheral change, like acquiring new vocabulary items that we forget about, etc. If the cognitive system has stabilized in state L, we'll say that it's Peter's cognitive system. If Peter's cognitive system stabilizes in the state L, we'll say that Peter knows/speaks/has the language L. To say that I have some special variety of that crazy mass of things called English is to say that my cognitive state stabilized in that particular system when I was ten years old or something. I'll come back to sharpening that up in a moment, but as a first approximation that's what it means to say that somebody has a language. The language, then, is just the state that the cognitive system reaches. That's all it is.

To avoid pointless terminological controversy as to whether this is really language or not, notice, incidentally, that we know in advance that the terms of common discourse are not going to survive into a theoretical discussion. Everybody knows this in the case of the natural sciences. Nobody cares whether the words 'energy', 'liquid', or 'momentum' as it is used in physics have the same meanings as the words in natural language, and of course it never will. It would be an astonishing miracle if a word of natural language could survive the transition to the natural sciences. There's no reason why that should happen. The words of natural language developed the way they developed – they're part of the natural world – and the constructions of the natural sciences are created by our science-forming faculties in an effort to come to terms with the world, and it would be astonishing if the same concept is carried over. Nobody expects to find a counterpart in the natural sciences to notions like heavens, earth, water, work, etc. There was a time when it was believed that there would be counterparts, but that's long in the past, and should be.

This is something that comes up in the discussion of contemporary theories of reference, so I'll come back to it in that connection.

Certainly, words of normal mentalistic discourse like 'language' are not going to survive the transition to a theoretical construction – the development of a theory in which we try to understand the way language is used and understood. When I say that that's what I mean by 'language', all I mean is that that is about as close a counterpart, within a considered theory of language, that we seem to be able to get to something like the intuitive notion.

It's interesting that in the natural sciences everybody just takes this for granted, but part of the irrational dualism of the mental sciences is that this is considered a big problem. People want to know if you have *really* captured the notion of a language when you characterize it that way. Nobody asks whether physics has *really* captured the notion of liquid, because nobody cares.

There was an article in *Science* (doi.org/10.1126/science.255.5051.1523) which was trying to decide whether a pile of sand is solid, liquid, gas, or some other form of matter. The point is, it didn't occur to the scientist, nor would it occur to anybody else, to go on asking people what 'liquid', 'solid' or 'gas' means. It doesn't matter what liquid means, what matters is a state of matter that developed in the natural sciences, and then you can ask whether a pile of sand is one or another state of matter (and it doesn't seem to be any of them for various reasons that I didn't understand). In the human sciences, these are considered problems. But we should overcome that. It's just of no interest. Since they are considered problems, let's make up a technical term so that we can avoid pointless debate.

I'll call this thing *I-language*, which is a technical term. The letter 'I' is used because of a fortunate accident of English. The concept that's developed happens to be a concept that is internalist, individual and intensional. To say that it is internalist means that it has to do with what's inside our heads, and not in the outside world. The concept is internalist in that sense. It's individual in the sense that it has to do with Peter and not some community to which Peter belongs (there are no meaningful communities). It's internalist and individual in the obvious sense, and it's intensional in the technical sense, which has to do with the generative procedure. When you have a generative procedure, like a computer algorithm that spins out an infinite number of things, you can look at that procedure or function in its *extension* or *intension*. If you look at it in extension, you're looking at the set of things it generates. If you look at it in intension, you're looking at exactly how it works.

Suppose you have some program for getting square roots – you put in 4 and it gives you 2 and so on. If you look at its extension, it is the set of pairs (4, 2), (9, 3), (25, 5), etc. If you look at the intension, it's the specific procedure that is used to get that result (and you could've used a lot of different procedures to get that result). For example, in the case of formal arithmetic, if you think of the axiom system as being the function that enumerates it, and if you look at this function in intension, then you care which axioms it exactly has (maybe it has Peano's axioms). If you look at this function in extension, you are looking at the set of theorems it generates, but with any set of axioms.

This approach is intensional. We're interested in the exact character of the generative procedure, and not in the set of things it generates (you are interested in those too, but that's on the side). Let's be clear about what the extension of the generative procedure is – the set of linguistic expressions, not the set of signals. And the extension is interesting, but it's derivative. What's important, at least from a natural science point of view, is how it's done – what the mechanisms are. Ultimately, you want to know what the brain mechanisms are. That means you're interested in the function in intension.

I stress this because it's completely opposite to the perspective taken in the reflective parts of the cognitive sciences and the philosophy of mind. What people are concerned with there is the extension. And as the extension, they don't pick the class of linguistic expressions – they pick another set which may be derived from the set of linguistic expressions or may not be. It's a set which is sometimes called a 'formal language', or a set of "well-formed expressions", or something like that – an equivalent of this, if you're looking at the generation of theorems rather than of expressions, would be the set of theorems in arithmetic. That's formal language.

There's no reason to believe that formal languages even exist, but if they do, they would be some kind of reflection of the actual extension by some further procedure. The actual extension itself is what the thing in your head (which has one or another form) grinds away. For example, suppose I'm a cognitive scientist interested in how people do multiplication, then what I'm interested in is the actual function that's going on in their heads. I'm not interested in the triples (X, Y, Z) such that X is the product of Y and Z. I knew that already. I'm interested in what procedure they use to get that. I'm interested in the function in intension, and the reason is because it's the function – the generative procedure – viewed in intension that is close to the mechanisms. In fact, the function in intension is an abstract expression of mechanisms. The function in extension is very remote from mechanisms.

Notice that the function in intension is also close to data – the data of experience. If you've learned a language, you can present it with some kind of data, and that data entered your mind and into your language faculty in its initial state (state 0), and various things happened, and out of that came your stable state. You reached a stable state using that data, and that stable state will include a particular generative procedure viewed in intension. That stable state is not an infinite set of linguistic expressions. Your mind is not an infinite set of linguistic expressions. It's not an infinite set of anything. It's a finite object, and the finite object that it is is the generative procedure (I-language) viewed intensionally. If there is an extension, viz., the class of linguistic expressions, then it's more remote from the data.

Again, this is the opposite of the way people look at it. The way people usually look at it is that they say that you have the data (the behaviour you observe), and then you somehow make an induction from that data to the set of linguistic expressions, and then you somehow find a characterization of them, which is a generative procedure, and then all kinds of debates go on about how you do it. Usually, what is said is that you go from the data to a formal language, but that doesn't make any sense at all because there might not be such a thing.

So, let's fix it up so that it at least makes some sense. We will say that you go from the data to the set of expressions (each of them some collection of properties), and then comes the problem of what's called finding a grammar (a grammar is just an I-language). So, finding a grammar means finding an I-language that enumerates and characterizes that set of expressions. And this step is the one that's considered the controversial one. The previous step is the one that's considered the straightforward one, namely, that it's some kind of induction (whatever that means).

If you look at the philosophical literature (say, David Lewis, or Quine, or anybody who's talked about this), they'll say that the problem of going from data to a class of linguistic expressions (they'll usually say 'a formal language') is sort of understandable, but to go from a set of linguistic expressions to a particular characterization of them rather than some other one – that's the part that's incomprehensible, mystical, etc. The truth of the matter must be the opposite. It just has to be.

There's no way of going from the data to the set of expressions. You can't go from finite data to an infinite formal language, or to an infinite anything. There's no way to go from finite amount of data to any infinite set without going by way of some finite characterization of that infinite set. You can't grasp an infinite set if you can't grasp a way of characterizing it. That means that if you've gone from data to

an infinite set, you've done it via a particular finite procedure that characterized the set. So, the first thing that must've hit the child in acquiring the language must've been the I-language, and then, of course, automatically he has the extension. Going from the data to the set of expressions is an impossible act. Induction (if such a thing exists) is a matter of going from finite data to a finite characterization of an infinite amount of data. It has to be that. It can't be to go from finite data to infinite data, because the infinite data doesn't mean anything to a person. It means something in set theory, but when we're talking about the natural world, there are no infinite sets in our heads. There can only be a finite characterization of it in our head. The usual debate about this that goes on has it completely backwards.

The problem of going from finite data to a finite characterization of infinite data is a straightforward one. It's the normal problem of biological development. It's no different from the problem of going from nutrition to a chicken. You have a germ cell that has some genetic endowment. It has a nutritional environment, which includes things like oxygen. Things happen later, and the final product is a chicken.

Credit: microscopy-uk.org.uk (Michelle Leung)

If you try to say that you don't end up with a chicken but with an infinite set of actions of a chicken – that you go from the nutritional input to the infinite set of actions of a chicken – that would be stupid. You have to get to the chicken first, then you can start talking about its infinite set of actions. Similarly, you must get to the I-language first before you get to anything that's done with the expressions that are formed.

Certainly, the performances, the manifestations, the speech acts, the signals, etc., exist. That much is uncontroversial within a naturalistic framework. It's just as uncontroversial here as it is in embryology. But, strikingly, in embryology, nobody ever debates it. There's no field of embryology, or philosophy of biology, or whatever where people ask, 'How do you go from nutrition to a chicken without first going through all the actions of a chicken?' I'm almost certain there's no field of philosophy which makes that proposal, but one might ask why not, because it's the same as the standard view in the study of language.

The standard view in the study of language is that the child goes from data that it observes to an infinite set of behaviours, and then comes this big problem of characterizing it, and then people get into debates about it that it's subject to a fatal indeterminacy or something. That would be like going from the chicken embryo's nutrition to the infinite set of chicken acts and then raising the philosophical question of how you can get to the chicken from the set of acts. That would be the counterpart of it in embryology. It's completely stupid and it's so stupid that nobody thinks about it. Even without knowing anything about embryology, it would never occur to anyone to ask that kind of a question. It's only in the study of mind that people ask such questions and come up with very crazy answers. Those are the examples of the kind of methodological dualism that I had in mind. I'll come back to them with specific references and quotes.

In any event, the naturalistic way is to say that you go from data to an I-language that is understood individually, internally, intensionally – to a specific characterization of the function, and not to some set that it enumerates. Having done this, the set of expressions that is determined by the I-language is a collection of instructions for the performance systems. The performance systems use those instructions to do things like articulating, or interpreting what you hear, or to talk about the weather, to express your thoughts, to ask a question, etc., but the I-language itself and the set of expressions it generates contains all the information that is used by the performance systems to do these things, and it accounts for all kinds of relationships.

Notice that it is quite possible in principle for two people – Peter and Mary – to have the same set of linguistic expressions but different I-languages, just as they could have a different way of computing square roots and come out with the same set of square roots. It may be that the nature of the language faculty rules this possibility out. It's probably true from what we now understand that the nature of the language faculty is so restrictive that it doesn't allow this possibility to be realized, but if so, that's just an interesting empirical discovery about what's called

Universal Grammar (UG) – the initial state of the language faculty. It would say that UG so restrictive that this possibility just can't arise, although it certainly did arise for arithmetical operations (people use different algorithms to carry out arithmetical computations), and we can't rule that out if it's not the case that it's a discovery (an interesting discovery about the near-uniqueness of languages) that there aren't many options open, that there aren't many I-languages around.

In these terms, one can account for many properties of sound and meaning, but let's be clear about how we're accounting for them. Let's take Peter and suppose that Peter has the I-language L, which is a particular generative procedure that forms an infinite set of linguistic expressions. Suppose I want to say for Peter that the word 'pin' rhymes with 'bin', that they have a formal relationship between them – rhyme. Or suppose I want to say that for Peter, "John killed Bill" entails "Someone is dead." It's reasonable to say that Peter's I-language expresses those relationships, but bear in mind what they are. These relationships, for one thing, are on a par – they're both relationships between linguistic expressions. 'Pin' has a certain class of properties and 'bin' has a certain class of properties, and to say that those two classes of properties rhyme is to say that there is a certain formal relationship between them (not a trivial relationship to describe, incidentally). To say that "John killed Bill" entails "Someone is dead" is to say that a formal relationship holds between these two classes of expressions – this formal relationship is also intricate to describe, and it seems that in order to describe it you have to look at decomposition of concepts, etc., which are a specific part of the properties of those expressions. But after having done so, you can express the relationship of 'entailment'.

In both the cases, these are formal relationships of pure syntax. It's just stating formal relationships between symbolic objects. What entitles us to call it 'rhyme' and 'entailment'? What entitles us to call them that is the way the I-language is embedded in the performance systems. The I-language is embedded in the performance systems in such a way that the formal relationship of rhyme comes out as something we perceive as rhyme in a phenomenal sense – and people do perceive rhyme. Very small children will pick out rhymes, and that's why you read them nursery rhymes, because they know what rhymes without any instruction.

So, rhyme is something very easily perceived, and we want to explain that, and the way we will explain it, since this is science, is by attributing some structure to the organism. And the right structure, as far as we know, is to say that there's an I-language that establishes formal relations between 'pin' and 'bin', and that it is embedded in the performance systems that operate in such a way that this set of linguistic expressions is understood by (say) my 2-year-old granddaughter as a

rhyme – we can carry out as many experiments as we like to determine that that's the way she's understanding rhyme.

Exactly the same is true about entailment. The relationship between "John killed Bill" and "Someone is dead" is a relationship of entailment (rather than something else) if the performance systems work in such a way that they interpret the formal relations that exist as a relation of truth preservation. Now, on the entailment side, the problem is harder to deal with. We have a lot more evidence about rhyme and we can make up easier experiments about rhyme than we can about entailment, but, basically, entailment is the same thing.

And if you really try to work out the rhyme business, it wouldn't be so trivial either, and I don't think anybody has ever done it. I'm not aware of any real theory of rhyme. We just sort of know it. And if you tried to work out the rhyme business, you'd naturally run into plenty of problems.

And nobody has paid much attention to the problem because, again, things on the sound side of language are not subject to this irrational dualism. Somehow on the sound side of language, we're able to be sane and treat questions the way you treat things that happen below the neck. It's when we get to the meaning side that all sorts of ideological issues start to arise, and irrational dualism starts to enter, and you get into big debates about the notion of entailment.

There are interesting problems about entailment, but, fundamentally, they don't seem any different than the problems about rhyme. The rhyme problem is not entirely understood, but it doesn't bother anybody because you know how to understand it better. The same seems to be true of entailment. We have plenty of reliable evidence about the semantic properties of expressions and semantic relations between them, and like any evidence it could be wrong (you never know if evidence is true, even in physics and certainly here), but we have fairly reliable evidence. You could replicate the evidence, sharpen it up, do statistical tests on it, and dress it up as much as you please. There are even some theories about entailment. The problems that remain are not different in principle from problems about rhyme, at least from a naturalistic point of view.

Notice that the I-language itself doesn't tell you whether these formal relationships are rhyme or entailment, or something else, or nothing – there are all kinds of formal relationships between sets of properties. In fact, you could imagine another creature which has an I-language just like mine but it's linked up to performance systems that guide locomotion. So, if the performance system produces

a seven-word sentence then the person walks in this direction, if it produces a twelve-word sentence then it walks in that direction, and if it produces a wh-phrase then it walks faster or something. It's perfectly possible to imagine an I-language that is a set of instructions for locomotion, in which case these formal relationships will still exist, but they won't be rhyme or entailment, they will be some other thing. The relationships are rhyme and entailment because of the way the whole language faculty is constructed. Remember, we're talking about a physical object here (there being no other kind of object), and as an object it has its own structure and properties. It appears to be the case, as we know, that its properties are the cognitive system embedded in performance systems, where the cognitive system yields infinite sets of constructions that have formal relations between them that are interpreted, acted upon, or manifested by the performance systems in ways that yield phenomena like rhyme, entailment, and so on. That looks like a true description of the language faculty and language use. The point to stress here is that no philosophical questions (for a naturalist at least) arise in the case of entailment that don't arise in the case of rhyme – all that arise are empirical questions of varying difficulties. This is roughly what a naturalistic picture yields, and we'd like to look at it in more detail later, but let's stop here and start looking at the general questions.

I should say that this is considered highly controversial (if not totally absurd) by people who think about these questions (not by the people who do it). It's almost universally thought of as either highly controversial or absurd – they're the domains of thought (mostly philosophy of mind) that are concerned with these issues.

The most general question is – is this the right way to proceed at all in the study of humans above the neck? That breaks down into a number of sub-questions, all of which are raised (not by me). So, we can ask various questions about it, and maybe they're right.

Well, what kind of questions can be asked? One question that can be asked is whether it's improper or controversial to describe the brain in this fashion, apart from questions of truth and falsity. Is there something improper about using such terms to describe the brain? Is it controversial that the brain is (say) a system that has states? – Apparently not. We describe every other system as something that has states and that doesn't seem to be controversial. Is it controversial to say that these states have properties? – Again, we do it all over the place, so that wouldn't seem controversial here. Is it controversial to say that one of these properties is being a generative procedure? – That can't be because this is one of the best understood properties there is. There are very few properties that are as well understood as this one. So, proceeding step by step, there doesn't seem to be anything controversial in

the moves that we've made. We're saying that the brain is implicated in the language faculty (apparently a part of the brain is). That part of the brain has various states and subsystems, and there's a proposal about them. Those subsystems have various states that they go through. The states have certain properties, like the property of being a generative procedure. We're all within really quite well understood notions. Nothing is as well understood in the natural sciences, so far at least. So, that part can't be controversial. If there is some problem, then it is lying somewhere else.

The second kind of question that arises is whether the results are correct. Is the story true? Does the brain have this architecture? Does it have these states and these subsystems and these properties? That's where the substantive issues arise, and we can put those aside for the moment. That's not the kind of objection that's being raised. What's raised is not, 'Look, you made a mistake. It's a different kind of generative procedure.' That's the kind of debate that goes on internal to the assumption that all this makes sense. These are where questions of truth and falsity arise, but they're not relevant here.

A third kind of question that arises is whether there's some alternative to looking at language and the use of language which would have better results, or a broader reach, or lead into other areas that are not reached this way, and so on. Anybody with a naturalistic temperament will, obviously, keep an open mind to that, and will expect that it's probably true. It's always been true in the past in all the sciences that everything is wrong. Therefore, it's probably true today that everything is wrong, and that we will be shown some better way to do it. But that fact in itself is not that interesting, because we are on a normal course of inquiry. The question really is – is there some alternative around that we can evaluate? And as far as I'm aware, there isn't. Therefore, it doesn't seem that there's anything to look at.

In fact, if we look at other approaches to language (like sociolinguistics, say) carefully, it always seems to presuppose all of this. They may not talk about it much but if you lay out carefully what they're doing, they always seem to be presupposing exactly this, and if you try to reconstruct it without presupposing this, you really reach absolute nonsense. That's true of even the people who deny it. People often deny what they're doing. It's not unusual. It seems to me if you see the actual work in sociolinguistics, anthropological linguistics, etc., you invariably find that all of this is being tacitly presupposed. I'm really not aware of an alternative way of proceeding that's been proposed. So, the third sub-query doesn't get you anywhere.

Now, there is a class of questions about whether the notion I-language comes close to the common-sense notion of language, and whether having an I-language

(having the cognitive state L where L is the I-language) comes close to *knowing* a language, *having* the knowledge, and so on. In my opinion, it comes pretty close, that is, the I-language is pretty close to at least one standard common-sense notion of language that commonly appears in the traditional literature – that language is a way of expressing your thoughts. To have a language is to have a particular way to express your thoughts. That notion is pretty close to I-language. As for having an I-language, that to me seems to come very close to having the knowledge.

This is very remote from philosophical theories of knowledge, which are usually phrased in terms of abilities, dispositions, relations between people and propositions, and so on, but those theories are just flat wrong. They have almost nothing to do with what we call having knowledge. So, not being remote from them is a problem. So, it is very remote from standard theories of knowledge.

But I don't think it's very remote from the notion of knowledge as it is actually used – like when we say that somebody *knows* the construction business, or if somebody *knows* his way around Girona, or something like that. That seems to me very similar to this notion of knowledge, though very remote from philosophical theories of knowledge, and certainly not expressible in terms of dispositions, propositions, etc. However, whether that's right or wrong is not very important, because it just doesn't matter a lot whether the concepts we come up with closely match those of common-sense discourse, any more than it matters in the case of 'energy' or 'undecidability' or 'angular velocity' or whatever you pick. You pick the concepts that you need to make sense of things, and if they're not close to those of ordinary discourse, so much the worse for ordinary discourse (which is fine for its own purposes but its unavailable for reflective, rational inquiry). So, that's not a real issue.

What I'd like to do next is to turn to various approaches that reject this entire enterprise and insist on what I would call a methodologically/epistemologically dualist approach, which is some kind of non-naturalistic dualism. I want to stress in advance that the advocates of the approaches I'm going to be discussing would never accept this characterization of their position. In fact, they regard themselves as hard-headed scientists, so they would completely reject what I'm saying about them. I'm going to try to argue that they're anti-naturalist (Quine, for example), and that everything that comes out of that tradition is a radical departure from naturalism. Obviously, Quine wouldn't admit that. He regards himself as a prototype naturalist and regards all this as mysticism. And I want to suggest that it's the opposite, and that it goes across the board, covering virtually all the reflective literature on the

philosophy of mind, cognitive science, AI, etc. That's the topic I want to turn to next.

Lecture #3

YESTERDAY, I OUTLINED what seems to me the proper course for a naturalistic theory of language. I didn't sketch any details, and I hope we can come back to that later, but it seems to me a model of what a naturalistic inquiry should be regarding any aspect of the phenomena that fall in the traditional domain of the mental. And I stress again that there is no metaphysical distinction between these phenomena and others, although there may well be an epistemological distinction – that these phenomena, to a crucial extent, may be beyond the cognitive reach of a particular creature, namely, humans. Mental phenomena might be a mystery to us – a conclusion that's not terribly surprising.

I also described the collapse of any naturalistic variety of dualism. For example, the Cartesian variety collapsed with Newton's demonstration that the ghostly properties of mind pervade the entire world – that matter, down to the level of elementary mechanics, also has ghostly properties from the point of view of the traditional mechanical philosophy. With that discovery, we abandon the common-sense view (at least for the purposes of theoretical inquiry) and move to an intellectual stage where we search for the best theoretical explanation, without any expectation that it will conform to our intuitive requirements – which simply are another phenomenon of nature that have to be explained.

I ended yesterday by saying that I would like to turn to approaches to the mental that reject entirely the naturalistic approach that I sketched yesterday and insist on a kind of radical dualism. I stress that the advocates of the positions that I'm describing would never accept this characterization of what they're doing. On the contrary, they consider themselves hard-headed scientists pursuing naturalistic inquiry in its most perfect and advanced character. But I will try to show that what they're adopting is a form of radical dualism, and a form that is completely irrational and untenable. Furthermore, I'll try to show that this is extremely pervasive – it includes virtually all of the more considered, reflective, thoughtful work on the cognitive sciences and on the philosophy of mind (excluding the actual experimental work, which goes on with such successes as it has).

Let's suppose that Peter and Mary have arrived at two different I-languages L1 and L2, respectively. A theoretical possibility I mentioned yesterday is that L1 and L2 could have the same extension, i.e., they generate the same set of linguistic

expressions. That possibility would arise from the fact that they had received somewhat different arrangements of data, and that data had led their common language faculty to slightly differ in conclusions, and which just happen to coincide in the linguistic expressions they form.

Now, it may be (and is very likely true) that in the real world that can't happen, because the language faculty is so restrictive that it excludes this possibility. But, that's an empirical issue, having to do with the narrowness of the language faculty and the limits on the kinds of languages that make possible. Conceptually, there's nothing surprising about this – it may very well turn out to be the case.

What I've just said is commonly regarded to be completely meaningless and absurd. There's a major current in modern philosophy which considers this comment simply as an absurdity, and the reason is that having the same linguistic expressions means having essentially the same behaviour. The way you speak, the way you understand, etc., depends on linguistic expressions (collections of phonetic and semantic properties). And if two people have the same collection of expressions, they're going to react the same way to what they hear and they're going to express themselves the same way, and over some large range of behaviour, they'll be indistinguishable. Hence, the theory has it that it is unintelligible to attribute to them two different states of mind, because the states of mind are just a characterization of the range of behaviour they exhibit.

Similarly, suppose that in the case of Mary, given a fixed set of linguistic expressions, I have a choice between assigning to her an I-language L2 or L3. As a scientist, I might not be certain from the data available to me whether to assign to Mary L2 or L3 (let's say for concreteness that they still generate the same set of linguistic expressions), and that would remain the case even if I had a theory of Universal Grammar (UG), i.e., a theory of the language faculty. Recall that the language faculty is this fixed species-specific system that permits Peter and Mary to take certain data and form in the mind an I-language.

Suppose I have two different theories of the linguistic faculty. One of them (T2) would lead me to conclude that Mary has L2 and the other (T3) would lead me to conclude that Mary has L3. Suppose I discover that there's somebody called Wong who speaks Chinese and has the same human brain that we have (since we know that there's no special genetic adaptation to learning Chinese). Suppose that in the case of Wong, T2 allows me to explain how Wong gets his knowledge from data, but T3 doesn't allow me to explain that. If I assume that the language faculty is described by T3 and then I look at Wong and his data and then I come out with

some I-language L4, that happens to be the wrong one (which I can tell by looking at Wong's behaviour). I can tell it's the wrong one since it generates the wrong expressions. On the other hand, if I pick T2 and apply it to Wong, then I get some other language, which is actually Chinese and comes out right.

Now, any natural scientist would tell us that for Mary we have to pick L2 and not L3, the reason being that L2 is consistent with a more general theory of the language faculty that explains Wong as well, whereas L3 is inconsistent with that theory and is only consistent with the theory of the language faculty which makes wrong predictions for Wong. The logic here is quite straightforward, and in the natural sciences would be totally uncontroversial.

However, we are instructed by philosophers that this is impermissible in the special case of language. We are not allowed to choose for Mary L2 over L3 on the basis of the fact that if we choose L2 we can also explain Wong. The language we assigned to Mary has to be based solely on Mary's behaviour and not on Wong's. Wong's behaviour is irrelevant in principle to attributing an I-language (often called a 'grammar' in the literature) to Mary. That's the instruction that is handed down by the contemporary philosophical tradition, which is very curious. The most prominent advocate of this position, the person who more or less established contemporary analytic philosophy, for which this is a crucial principle, is W.V.O. Quine, and he actually takes a still stronger position.

Remember that there is another notion here called 'formal language', which is a class of well-formed expressions, and there are some problems about this – nobody knows what it is. But let's put those aside and pretend that we know what the formal language is.

Now, clearly, you can have different classes of expressions that coincide in their signals – they coincide in their signals but differ in all sorts other respects. For example, they could differ in how they associate signals with semantic properties, and so on. So, you could have many different sets of expressions which coincide with the same set of well-formed expressions, pretending we know what the set is.

Now, according to Quine's version, if Mary and Peter correspond in their formal languages (in the signals that they accept), then it makes no sense to say that one rather than another I-language is true of Mary and Peter, and we are certainly not allowed to assign different ones to Mary and Peter. Similarly, *a fortiori* we're not allowed to consider anything about Wong. As Quine puts it, if two grammars (I-languages) are extensionally equivalent, meaning if they're the same with regard to

formal language, then it is meaningless to claim that one rather than another grammar is in the mind of the speaker, and *a fortiori* it is methodologically wrong to appeal to the fact that one decision will allow us to explain Wong whereas the other decision will leave Wong unexplained – we're not allowed to do that either, because only the class of formal expressions is the database for the linguist in assigning a grammar.

That's the theory, and it has a kind of motivation – the child learning the language only hears signals and is picking up the I-language from the signals. From that Quine concludes that it is meaningless for the linguist (who he identifies with the child, which is obviously a mistake) to assign different I-languages on the basis of the same set of signals.

This is what is called the 'radical translation model'. Other followers of Quine, like Donald Davidson, go even further (while accepting this much). Davidson argues that it is a mistake to claim that what underlies behaviour is some specific set of mechanisms. As he puts it, it adds nothing to the theory of meaning (that's just the philosopher's way of talking about the theory of language) to assume that some mechanisms correspond to the theory.

For example, if one could find certain cellular mechanisms that are consistent with T1 but not with T2, that would be irrelevant in principle to the linguist. If I'm trying to determine Peter's I-language, all I'm allowed to look at is the set of linguistic expressions (signals) that Peter accepts. I'm not allowed to go on to do what any scientist would do and ask whether one theory of the initial state is consistent with the biological mechanisms and another one isn't. I'm not allowed to ask whether one theory of the initial state is consistent with Wong or Ahmed or Luigi or a different speaker. Those moves are barred to the linguist. The linguist is not permitted to study these aspects of the brain by the methods of the natural sciences. That's the theory.

Well, then, presumably, somebody is allowed to study these aspects of the brain. The injunction doesn't seem to be that no one is allowed to study this. Presumably, somebody is allowed to study the initial state and to make use of the fact that one theory of it accounts for Wong and another one doesn't. Somebody is allowed to study that but not the linguist, by the philosophers' stipulation. Since this is a purely terminological stipulation, let's just accept it – linguists are people who, by stipulation, are committed to total irrationality. They're not allowed to study any of the topics that anybody would investigate if they want to understand the nature of language. Now, having accepted this terminological stipulation, the rational move

is to completely abandon the ridiculous pursuit that is now called linguistics and to turn to this other subject in which we are allowed to look at Wong when we try to decide what's the right I-language for Peter, and to look at biological phenomenon if we can find any, and to do anything that a normal scientist would do.

In fact, this happens to be the actual practice of linguists – which is condemned in this tradition of philosophy, and which, in a final irony, prides itself on its naturalism and its adherence to the methods of the sciences. This seems to me a case of quite radical dualism, and I stress that this is a highly influential tradition. It dominates a large part of the contemporary philosophy of language and mind and feeds over into the cognitive sciences. It seems to me a pure irrational stipulation that in one building in the university you're not allowed to study the brain for some reason, and you have to go to some other building to study it, and here you are allowed to do exactly what linguists are now doing, but they're not allowed to call themselves linguists. This is what it comes down to. It is rather reminiscent of a comment that Voltaire once made about metaphysics, that it's a game with extremely elegant moves where you always end up where you started. That's more or less accurate in this case.

Now, Quine himself has a further extension of this curious theory. Remember, the linguist is restricted to looking at the signals, which are the well-formed expressions. And Quine should go on (I don't know why he wouldn't), at least, to say that the linguist is permitted to look at the way signals are associated with interpretations in a particular community, but he's not allowed to go any further than that. However, Quine in his more recent work has been willing to accept certain other evidence that would be relevant. One problem that he discusses is the problem of how to assign phrase structure. If you have a sentence "John saw Bill", there is a question of whether it is three or two different phrases or whether it has some other arrangement. Those are empirical questions that one might ask. Quine has argued that there's no truth to this the matter with regard to this distinction – the phrase structure is anything you decide – and the reason for that is that you get the same set of signals no matter which way you do it (which is consistent with his terminological stipulation).

However, more recently, he's willing to accept the idea that some other kinds of evidence might lead you one way or the other. The other kind of evidence is what's called 'psychological evidence'. Psychological evidence is distinguished throughout the tradition, including the cognitive sciences, from linguistic evidence. So, linguistic evidence, for example, would be evidence about how people understand referential relations – it turns out, for example, that if you say, "He thinks

John is intelligent", then in English, and in every other language, we know that the person who utters this sentence intends 'he' to refer to someone other than John. On the other hand, if you say, "His mother thinks John is intelligent", then we don't know whether the speaker is referring to John or to somebody else when he uses 'his'. Evidence of this kind is called linguistic evidence, and there's a huge amount of it. It turns out that if you use evidence of this kind across many different languages, you get quite strong arguments telling you that the phrase structure is a certain way.

Then there's another class of evidence called psychological evidence. An example of psychological experiments bearing on this matter is a set of experiments that were started by Tom Bever about twenty years ago, which refer to the perceptual displacement of noises. Suppose you give a person a sentence (a signal), and the person is listening to it, and in the course of the signal you introduce a click somewhere, and later you ask the person where they heard the click. What they will tend to do is to displace the click to the phrase boundary. Quine has agreed that this kind of evidence the linguist is allowed to use. He is allowed to use psychological evidence to determine where the phrase boundary is, but he is not allowed to use linguistic evidence (say) about referential relations and tons of other evidence of that kind. Now, that too is a very curious move.

Firstly, evidence doesn't come divided into categories. There's no such thing as linguistic evidence and psychological evidence. Evidence is just data, and data becomes evidence when you can interpret it within some sort of a theory, or else it is just random data. Hence the distinction doesn't make sense in the first place. But even granting the distinction, the story is backwards. The only significance of the click studies is that they correlate with the conclusions of the linguists. The so-called linguistic evidence about phrase boundary is vastly more convincing on scientific grounds than the perceptual evidence is. In fact, the perceptual evidence tells you nothing. The only reason why people like Bever and Fodor have interpreted the psychological data as supporting the phrase boundary between (say) the subject and the predicate is because that is what the linguistic evidence shows, which is really quite persuasive. Furthermore, the linguistic evidence fits into a theory (so it has theoretical consequences), which means there's a whole mass of indirect evidence that supports it, including evidence from (say) Chinese, evidence from historical changes, evidence from child language, etc. The click evidence has nothing supporting it at all. It is just an observation, an interesting piece of data, which becomes evidence because you can interpret it in terms of the much more firmly grounded linguistic theories. This is from a naturalistic point of view. But Quine takes the opposite view. His view is that you're allowed to appeal to the

psychological evidence in determining the phrase boundary, although it tells you absolutely nothing. Again, these moves are completely baffling from a naturalistic point of view. They make all kinds of distinctions that don't exist. They put matters on their head. They interpret the powerful evidence as weak evidence and vice-versa. They make no sense at all.

Now, we can work out a kind of psychoanalysis and figure out where this comes from. It comes from a picture of science as something that's done by people who wear white coats, possess equipment, do statistics, etc., science isn't what people do when they sit around and think about the structure of sentences and ask their friends what it means. But that picture is, of course, just nonsense. The linguistic experiments on (say) referential dependence could also be dressed up with white coats and statistics and equipment and so on, but nobody bothers to do it for reasons well understood in the natural sciences. You never do experiments beyond the level of precision that's relevant to the questions you're asking, and the questions about phrase structure are well enough answered by the very reliable judgments about sentences like this. And if you want, you could dress them up as more complicated experiments, but that would be as if the physicist was to carry out experiments to the 10th decimal place when he doesn't even understand the 2nd one. That would be pointless. You don't use more experimental precision than bears on the topic at hand.

There may come a time (and maybe it has already come) when linguists ought to go beyond this level of precision and think about sharper experiments that separate out things that we may be mixing up, in which case you do better experiments. But there's no break between the whitecoat-wearing psychologist who's doing it with a tachistoscope and the linguist who is just asking their friends. It's all just ways of getting data that you can use as evidence.

I presume that that's the origin of the distinction, but whatever it is, it's a kind of very radical dualism and an extremely sharp departure from the natural sciences. It's hard to imagine, in fact, a sharper departure. And, again, this is accepted across a broad range as meaningful. Throughout the cognitive sciences, there is a lot of material that claims that if linguists really want to establish their theories, then they have to get psychological evidence, and that the linguistic evidence won't do. So, they'll have to get evidence from click studies, which are completely meaningless, whose only significance is that there's a way of interpreting them so that they correlate with the linguistic evidence.

Again, these are very strange moves from a naturalistic point of view, but they are intelligible in terms of a kind of radical methodological dualism which states that you're just not allowed to study humans the way you study every other creature in the world. That's the stipulation.

Another issue that arises in this connection (and that's again all over the place) is the matter of access to consciousness. This plays a big role in philosophy of mind and cognitive science. One place where this shows up is in a distinction that's supposed to exist between psychological hypotheses and philosophical explanations. The discussion is usually framed in terms of problems in the theory of meaning, but that's only for historical reasons. To the extent that the arguments make any sense, they would apply to the study of sound, the study of syntax, etc., but it just happens that the tradition out of which they come has largely been concerned with meaning rather than (say) phonetics.

Take another very influential contemporary philosopher, Michael Dummett. He's one of the major modern philosophers of language and mind. Dummett has argued repeatedly over the years that we have to make a crucial distinction in epistemology between psychological hypotheses and philosophical explanations. And here's how his argument works. Suppose that the naturalistic approach is so successful that it is able to give a complete naturalistic account of what happens when a signal strikes the ear and gets computed through the perceptual system and the brain feeds into a theory of action and leads to a certain action. Suppose we have a theory that accounts for that whole chain. Suppose further that we solved the unification problem, that is, we've shown how this theory of what the brain is doing relates to an account of cellular activities and chemical interactions between them, etc. In that case, Dummett tells us, we would have a psychological hypothesis, that is, the theory that we constructed would be a psychological hypothesis about people, but it would not be a philosophical explanation, and the reason is because "it would not tell us the form in which the body of knowledge is delivered".

Remember that one part of this theory is inside Peter's brain – a cognitive system (I-language) which stores knowledge, and is accessed by the performance systems and the perceptual apparatus and the theory of action and so on.

Dummett's point is, even though you had a perfect theory of this by scientific standards, it would not tell us the form in which the knowledge is delivered. The idea is that in Peter's I-language (in the cognitive system), there is knowledge stored, and Peter's mind accesses that knowledge to hear, to act, etc., but we don't know the form in which the knowledge is delivered to Peter. It's only if we knew the form in

which the knowledge is delivered that we would have a philosophical explanation, not merely a psychological hypothesis. And the argument is framed particularly in connection with the theory of meaning, which is what Dummett, being a philosopher is naturally interested in.

Suppose my theory gave a complete scientific explanation that explains how Peter understands sentences like "He thinks John is intelligent" and "His mother thinks John is intelligent", and is, therefore, a theory of meaning. According to Dummett, that complete theory would be a psychological hypothesis about Peter's knowledge of meaning but not a philosophical explanation, and it would not justify us in attributing knowledge to Peter, i.e., we would not be able to say that Peter *knows* the rules that lead to this conclusion.

There is a part of linguistic theory called binding theory, and it has various principles that predict that things ought to come out like this across all languages (we're now pretending that everything we believe is true, and even that we've gotten way beyond what we currently understand, but that's permitted in this intellectual exercise).

Pretending that binding theory is true and it leads to the conclusion that for Peter these two sentences are interpreted exactly the way he interprets them, we are still not permitted to attribute to Peter knowledge of binding theory. We are not allowed to say that Peter *knows* the principles of binding theory, although binding theory is part of his cognitive system and is accessed when he acts. Obviously, when you act you access your knowledge or your beliefs, but we're not allowed to say that in the case of Peter. As Dummett puts it, the reason why we don't know the form in which the body of knowledge is delivered is that

Peter can't tell us that he is following the principles of binding theory – we have this kind of evidence about Peter, but if I ask Peter whether he is following condition C of the binding theory, he cannot answer. Even if Peter is a linguist and knows what condition C is, he can't introspect and say that that's the computation that's going on in his head. You can't answer that any more than you can answer the question of how you digest your food. You couldn't introspect and say that you do it the way the biologists tell us we do. Or if somebody asks you how you perceive a triangle, you can't introspect and say that you do it the way the visual psychologists claim you do. Binding theory is just like that – not available to introspection. Therefore, we don't know the form in which the body of knowledge is delivered, and therefore we don't have a philosophical explanation.

Notice that for science, the account that we're talking about tells us everything that can possibly be asked about the form in which the knowledge is delivered (by assumption). Our account has answered every question that science can raise, but it hasn't crossed some bridge to philosophical explanation.

This again seems like a paradigm example of some kind of philosophical dualism. In the case of digestion, visual perception, or the motion of the planets, we don't insist on a philosophical explanation. That goes beyond what science can deliver. That would be a joke if you did. There was a time when people used to insist on that, but it's long gone. We now say that what science can deliver ends the story. But in the case of humans above the neck (in the case of language, for example) we're not allowed to do that.

Now, presumably the same argument would hold for sound (though Dummett doesn't say so). If I conclude that Peter judges that 'pin' rhymes with 'bin', and I make up a perfect theory that explains this, that would not tell me the form in which the knowledge is delivered. I wouldn't be able to attribute to Peter knowledge of the principles of phonology that led to that consequence. No matter how well-confirmed the theory was and no matter how well it was unified with cellular theories, I still would have only a psychological explanation. That's a very curious position, and to extend its oddity let's push it a little further.

Suppose there's a Martian who happens to be exactly like us. He has the same brain as us, and the same theory of our brain applies to his brain. But there's only one difference – to the Martian, the rules of binding theory are accessible to consciousness (and in fact all the rules of his I-language are accessible to consciousness). He is exactly like us, but he has an inner eye that inspects his internal computations. Now, if we give the Martian these sentences and he makes the same judgments that we do (by assumption), and then we ask him if he has applied condition C of the binding theory and he says, 'Yes, that's exactly how I did it.' Just as if somebody asked you how you multiplied two numbers and you might say that you did it in this particular way. Similarly, if the Martian says 'pin' rhymes with 'bin', and if I ask him if he followed the rules of metrical phonology, he can introspect and confirm that those are the rules that he followed. And let's assume he's not lying. So, this Martian can see the computations just as we can introspect about the computations when we carry out a particular algorithm (rather than a different one) to solve an arithmetical problem. That's the only difference between the Martian and us.

Notice that for the Martian, we now have a philosophical explanation of the form in which the knowledge is delivered – we have crossed this mysterious bridge, because the Martian has access to consciousness. We know nothing of any relevance about the Martian beyond what we know about the human, except that the Martian has this inner eye that we don't have that enables him to inspect the computations, but we're allowed to attribute knowledge to the Martian and we're not allowed to attribute it to humans.

Well, again, this simply departs radically from any scientific approach. As far as the sciences are concerned, the Martian and the human behave exactly the same way, and the Martian happens to have access to computations that the human doesn't. The question of consciousness is a fine one, but it has nothing to do with the problem of whether Peter and the Martian are following the same rules, or have the same knowledge, or whatever. It's just irrelevant. Nevertheless, it is considered critical.

The same distinction shows up in what's called 'first-person authority.' The idea is that when I say that in "His mother thinks that John is intelligent" 'his' could be referring to John, I say that with first-person authority. I have supreme authority over that judgment and nobody can question it. I have special privileged knowledge of that fact because I have first-person authority when I look at the room and see a bunch of people. Similarly with the Martian, he has first-person authority when he makes the same judgments. But according to this theory, we have an explanation for first-person authority in the case of the Martian, but in the case of the human, the theory that I have given (a perfect theory by assumption) leaves the question of Peter's first-person authority a total mystery, because Peter has no access to consciousness.

As far as I can see, from a naturalistic point of view, that's the wrong formulation. The theory of Peter doesn't make first-person authority a total mystery, it leaves the mystery exactly where it was before, which is the mystery of consciousness. That's as much a mystery after this theory was completed as it was before. But as far as first-person authority is concerned, there's no mystery whatsoever – Peter understands the sentences because that's the way Peter is constructed. There's nothing more to say. You're constructed in a way so that you get that interpretation, and that's a full account of first-person authority. There are no further questions to be asked, at least within a naturalistic framework. If something's missing, it must be because of some non-naturalistic requirement, and it would be important to bring out what that requirement is.

This discussion is usually framed in slightly different terminology. In this case, the terminology goes back, again, to Quine, and to Wittgenstein and others as well. It is the problem of following rules. If Peter is using condition C of the binding theory in drawing that conclusion, the linguist would like to say that Peter is following the rule 'condition C', which is how Peter draws the conclusion that these sentences have the interpretation they have. Now, that's supposed to be a big problem. There's supposed to be a mystery about the notion of following rules. Quine has suggested that we resolve the mystery by making a distinction between two notions of rule-following. He suggests that we distinguish between *fitting the rules* and *being guided by rules* ('Rule' here is a very broad notion. It includes, for example, a law of nature, a principle of grammar, a rule of chess, etc).

How does fitting the rules work? Let's take Kepler's laws of planetary motion. In that case, Quine says that we should say that the planets fit Kepler's laws but aren't guided by them. The planet doesn't think, 'I'd like to follow Kepler's laws, therefore I'll move over here', rather the planet just acts in such a way that it fits the laws. On the other hand, a person is guided by the rules (and this is only people) if the person consciously explains that that is what he is doing. For example, if I'm driving a car and there's a red light and I stop, and somebody asks me what I did, and I think about it and say, 'I followed the rule that says you're supposed to stop at a red light'. If this description is true, then we can say that the person is guided by the rule. So, those are the two categories: you're guided by rules when you can say truly that those are the rules you're following, and otherwise you just fit the rules, nothing else.

Notice that this excludes Peter. The scientist would like to say that Peter is following the rules of binding theory. Peter certainly isn't guided by them, because he can't tell you that that's what he's doing. The Martian is guided by them because he can say so. All we're allowed to say in Peter's case is that he fits the rules, the way the planets fit Kepler's laws. We're not allowed to say that these rules are part of Peter's cognitive system and that he accesses them and so on. In other words, this perfect theory that we pretended to have has to be thrown out because it goes beyond fitting– it actually attributes the rules to Peter. According to Quine's notion, no matter how successful that theory is (and it can achieve perfect success from a scientific point of view), we have to throw it out, and all we're allowed to say about Peter is that he follows the rules of binding theory, but they're not part of his constitution. That's the stipulation.

Now, here too there have been some very curious moves. There are even some curious moves about fitting rules. Take the example of Kepler's laws. The physicist

doesn't just say that the planets fit Kepler's laws. What he does is he attributes to the planets a certain property, namely, 'mass', and then he states certain laws, say, Newton's laws, and then it follows that because planets have mass, given Newton's laws they are going to behave in accordance with Kepler's laws. That's the actual step that's made.

So, if we want to say that Peter just fits the laws of binding theory, we have to do what the physicist does and attribute to Peter some property from which it will follow by the laws of nature that Peter can draw those conclusions. That would be treating Peter like the planets. Well, what property do we attribute to Peter? We could attribute to him mass, but nothing is going to follow from that. That's not enough. We have to attribute to Peter richer properties than that. For the scientist, the properties that you would attribute to Peter are the properties stated by the structure of binding theory – you would attribute to Peter a property of having a brain with a language faculty, and a cognitive system which includes binding theory, and a performance system that accesses it, etc. That whole complex of material is the property that I'm attributing to Peter, just as I attribute mass to the planets. If some thing up there in the sky didn't have mass, it wouldn't follow Kepler's laws, and if Peter didn't have this structure, he wouldn't follow the laws of nature and yield these conclusions.

However, we're not allowed to make that move in the case of Peter, by stipulation, which is another form of radical dualism. In fact, when people talk about following rules, it's within the category of fitting the rules, but in which you're allowed to attribute to the person properties that would account for why they fit the rules. And by stipulation the linguist and the psychologist are not allowed to do this (maybe the psychologist is, but not the linguist). Again, that's a pure dualist stipulation. There's a huge literature following this distinction, and the person who's developed this the furthest is John Searle.

Searle insists strongly, as most philosophers and cognitive scientists do, that access to consciousness is the crucial criterion for attributing rule-following. Wittgensteinians more or less do that too, although it is a little more qualified. But Searle explicitly demands that it is nonsense to speak of Peter following the rules unless Peter can tell you that's what he's doing. So, we're not allowed to do what the scientist would do in this case.

Searle, however, has recognized that there are certain paradoxes that follow from this insistence. One thing that's bothered him is the phenomenon that's called 'blindsight' in the psychological literature. It was discovered some years ago that if

a person suffers certain kinds of damage to the visual cortex, then in some regions of the visual field the person has no awareness of seeing anything. It looks blank. On the other hand, in some cases, it turns out that you can show experimentally that the person really is seeing something. For example, if you draw a cross or a circle in their blindsight region and you ask them if they see anything, they will say no, but then if you ask them to guess if it is a cross or a circle, the person makes the right choice – they'll say it's a cross if it's a cross and circle if it's a circle.

If you put in one part of their blindsight region a picture of a house on fire and in another part a picture of a house not on fire, and if you ask them what they see, they'll say that they see nothing, but if you ask them to extend their hand to some place where it won't get burned, then the person will extend their hand to the picture of the house not on fire. That's the phenomenon of blindsight.

Well, here there's obviously a kind of problem arising for this rule-following business, because if you take Quine seriously, you have to say that the person with blindsight has every reason to believe that he is perceiving the cross, the circle, and the house on fire just the way he did before. He just doesn't have any awareness of it. Quine, Dummett, and others will be forced to say that the person is no longer following rules, even though he is doing exactly what he did before. This, Searle recognizes, would be kind of paradoxical, but he doesn't want to draw that conclusion, so he introduces a distinction between what he calls 'blockage' and 'inaccessibility in principle.' In the case of the man with blindsight, he only has blockage. He doesn't have inaccessibility in principle to what he is perceiving – it's just some interference, some blockage. On the other hand, in the case of Peter following the rules of the binding theory, that's inaccessibility in principle.

Let's pursue this argument further. Let's go back to the Martian who talks English and has complete access to all the rules. Remember, the Martian does have accessibility to the rules. Suppose the Martian has a brain injury, and now he's just like the person with blindsight. The Martian keeps talking exactly the way he talked before (which, by assumption, is the same way we do), he acts just like before, but now he's lost the inner eye. The brain injury has eliminated his ability to access the computations. According to Searle, the Martian has blockage and not inaccessibility in principle. Just as the person with brain damage lost access to awareness but is acting as before and therefore only has blockage, the Martian, who, through a comparable brain injury, has lost access, only has blockage and not inaccessibility in principle. So, the Martian is still following the rules just as the guy with blindsight is still seeing, but without awareness.

Notice that the Martian is now indistinguishable from us. In fact, his brain injury might have turned the Martian into a human. Suppose the Martian had some little component up there that was making it possible for him to inspect his computations and conclude that it is binding theory, and suppose an operation just cut that component out. Now, the Martian is indistinguishable from humans but only has blockage and is still following the rules, while the human is not following the rules because of inaccessibility in principle.

Now, let's carry this thought experiment a step further. Suppose that among Martians there's somebody born with a certain genetic defect, making the guy indistinguishable from the Martian with the brain injury. This genetic defect of his just didn't allow that component of the brain to grow. This other Martian is now identical to the Martian with blockage, and identical to Peter, but it wasn't caused by a brain injury – it was caused by the genes. Now, we have to decide what we're going to say in this case, and, unless we want to return to absurdity, we would have to say that the Martian with the genetic defect still just has blockage – he's exactly like the other Martians, but something went wrong with his brain, even though he's born identical to humans.

Suppose this genetic defect is perpetuated for thousands of years, and now there's a whole race of Martians all of whom have the same genetic defect. Suppose they've now come down to earth and intermingled with us, and since they're identical with us we can't tell any difference. We would still have to say by this reasoning that these Martians only have blockage, while the rest of the humans (who have some other evolutionary origin) have inaccessibility in principle, and the Martians among us are following rules but the humans are not. In fact, that could be the reality. Maybe there was an invasion 50,000 years ago of Martians with genetic defects, and now half of you are descended from Martians and half from humans, and some of you are following rules and some are not, and we have no way of telling. That's what we are forced to, if we pursue this line of reasoning. It becomes obvious that something's gone wrong. In fact, to stop this course of absurdity from proceeding, we would have to cut it off at the first step and say that there couldn't be a Martian who is just like Peter except that he has access to consciousness. If we allow the possibility of such a species, we go off into this absurdity. Therefore, in order to block the absurdity, we have to say that there couldn't be such a species. Well, we now have the philosopher making a very curious empirical assumption about the possibility of a certain species. The philosopher is now driven to an empirical assumption that goes way beyond any biological knowledge, which says that a certain kind of species are impossible according to the laws of nature. When philosophy is driven to a crazy empirical assumption, you know that something's

very wrong. And tracing it back to the origins, what's wrong is the whole idea that access to consciousness has any relation whatsoever to attributing rule-following. Therefore, we just go back to the naturalistic interpretation – access to consciousness is an interesting question, but it just has nothing to do with rule-following.

In this post-Newtonian naturalistic approach, we just look for the best theory, and if the best theory tells us that Peter is following rules, then so be it. We have no higher criterion to which we can appeal. In the post-Newtonian era, there is no privileged category of evidence like, say, linguistic evidence or psychological evidence. There's no notion of philosophical explanation that goes beyond psychological hypothesis. There's no notion of, say, blockage vs. inaccessibility in principle or any such mysterious inventions which are introduced in an effort (which is certain to be a vain effort) to keep the whole thing from collapsing into absurdity.

I should say that these questions have been raised to John Searle in particular, and I don't see any answers to them. It seems to me that he kind of skirts the questions. It seems to me that unless he can answer these questions and others like it, the whole project totally collapses. And with it the whole Quine project collapses, which is based on the distinction between guiding and fitting, both of which are misinterpreted. Fitting is misinterpreted because it fails to notice that the planets fit the rules by virtue of a property that they have, and guiding is misinterpreted because it is far too narrow – it's based on the arbitrary stipulation that access to consciousness is a critical criterion (which makes no sense in the first place).

In the case of the Wittgensteinian variant of this (or Saul Kripke's) there's a somewhat different view of the matter. In the Wittgensteinian version, there's also a big discussion about following rules, but the examples are typically the rules of arithmetic. When we see somebody multiply numbers, how do we know that the person is following the rule of multiplication? How do we know that when you get to 5 times 25, it's not all of a sudden going to be 12862? After all, that's a rule too. There's a perfectly good rule of arithmetic that says – multiply the normal way up to big numbers, and from that point on make all the products equal to (say) pi. How do we know that the person is not following that rule, which is a perfectly good one? As Wittgenstein points out, no matter how much evidence you have, you will never know that. You might have the hypothesis that when you get to 100, all the products go up to pi. And when you get that far, you'll find you were wrong. But then you can change the rule and say that it's when you get to a 1000 after which they're wrong. For every mass of evidence that you accumulate, you're always going to have these further hypotheses. Therefore, Wittgenstein concludes that you can never

say a person is following the rules. It's kind of like a paradox. And Kripke, in a recent book, develops this much further.

Whatever the significance of any of this might be, we can put it aside, because the rules are not relevant. These rules are rules of a totally different category. The rules of binding theory (what Peter is following) are laws of the natural world. They're like Kepler's laws or Newton's laws. It is a law of the natural world (as it is claimed) that if you have a brain organized in terms of the principles of binding theory embedded in a performance system of a certain type, you'll understand sentences in a certain way. And questions do arise about the validity of natural laws, but those are just normal questions of inductive uncertainty. And as I said at the beginning, general questions that arise for physics don't have to trouble us when we're doing psychology. If you want answers to those questions, go to the physicists. There's no point in raising for psychology or linguistics exactly the kinds of questions of (say) inductive uncertainty, or accuracy, or legitimacy of theoretical postulates, etc., that arise for the most advanced sciences. And they arise necessarily, and they can't be overcome. They are just an inherent property of natural inquiry. It was well understood by the 17th century (after the Cartesian crisis of skepticism) by people like Gassendi that the problems of induction are inherently insoluble – all we can do is realize that we have some method of gaining better and better understanding while knowing that it can't have any absolute grounding. That's the result of the skeptical crisis and that's basically the position of the natural scientists, and it's a fair one for us to adopt too.

As to the specific questions that arise in the case of arithmetical rules, they are irrelevant, because the arithmetical rules are stipulated. Those are rules of a totally different type. We stipulate that 'plus' is going to mean so and so, and if you don't follow that, you are just not following my stipulation. That's like the rule of stopping on a red light or something. It's a stipulation. We can't be sure that people are following stipulations, but who cares. It's not relevant. Therefore, the Wittgenstein-Kripke kind of consideration can be put aside. We're simply concerned with rules that are of the type of natural laws – rules that Peter follows because of the way he is constituted.

We have to be careful here to distinguish two ways in which people can follow rules. For example, I know that the word 'chair' can be used to refer to a chair but not a table. That's following a rule. Now, I could consciously choose to violate those rules in a certain sense. I could use 'chair' to refer to tables. We might make up a code in which we want to confuse spies. We can certainly do that. In that case, we wouldn't be following the rules. And, of course, we would know that we're not

following the rules because we can't help knowing that 'chair' means chair, but we could just choose to overlook that fact and use these physical objects in some other fashion. That's a distinction of rule-following that one has to bear mind, but it plainly has no bearing on this question.

I want to turn next to further examples of this dualism and continue to try to show that it indeed is all pervasive.

Lecture #4

LET ME BEGIN by recalling the general plan. I'm trying to, first, defend a naturalistic approach to problems of language and mind. Second, I want to try to show that the most influential currents of contemporary thought that deal with these issues, primarily philosophy of language and mind but also spilling over to cognitive science, are both non-naturalist and radically dualist. I tried to show yesterday that those are two different things, that you could have naturalist dualism (like the Cartesians). But these are dualists in the non-naturalist way – methodological/epistemological dualists. And the third point I want to make is that all of this is a very serious regression since the 17th century and it should be completely abandoned without a trace being left, to put it as strongly as possible.

This morning, I gave some examples of radical methodological dualism with regard to (roughly speaking) philosophical issues – issues having to do with attribution of knowledge, the propriety of attributing rule-following, etc. I think the most charitable conclusion that can be drawn is that in the major thinking about these issues, we have a series of arbitrary stipulations about certain invented disciplines, one of them called linguistics and the other called philosophy. Linguistics is instructed by these doctrines that if it wants to study Peter's mind, then it must consider only Peter's behaviour (and maybe people in Peter's community), but it is not allowed to study Wong's behaviour and (say) cells. Those steps are ruled out, although any scientist would take them at once.

If we agree that that's what linguistics is, then the response is simple – just abandon it. It's a ridiculous pursuit and therefore abandon it and turn to this other study in which you're allowed to bring in evidence of whatever sort is relevant to understanding Peter's mind – Wong's mind, things about cells, aphasia, or whatever else. The answer to the first stipulation is simple – abandon this ridiculous vocation and turn to a serious topic, which happens to be the topic everybody is pursuing anyway, so it's not going to be a big change.

With regard to philosophy, it seems to be identified as a discipline that is concerned with a philosophical explanation different from scientific explanation, and is interested in something called attribution of knowledge and rule-following in a technically invented sense (the sense in which these things turn on access to consciousness). Now, it's possible to invent such a discipline (you can invent

anything), but that discipline has no connection to inquiry into the nature of the world. It has no connection with ordinary usage, which doesn't follow those principles. And as far as I can see, it has no connection with anything. It has a historical tradition, meaning one can see how it got there, but, again, it seems to me a pursuit that has no redeeming virtues and should simply be abandoned. That's my conclusion about the second stipulation.

What I want to do now is begin to turn step by step towards cognitive science, from what are considered philosophical issues, which just seem to be matters of terminological stipulation, to more substantive, empirical issues.

One substantive issue has to do with the growth and development of cognitive capacities. We know that they change. They're not the same at age 0 as they are at age 12. Therefore, some change takes place in our cognitive capacities – some kind of growth (or what's called learning, in my opinion, rather misleadingly) takes place. Here we, at once, run into a kind of dualism of a curious kind, so let's look at it.

There's a doctrine that is called 'innatism', and there's a big debate about it. As in the couple of cases I mentioned yesterday, this debate is one-sided. Lots of people write articles denouncing innatism, but nobody debates it. I'm supposed to be the main criminal here – the one who is mainly guilty of the crime of innatism – but I never really responded to any criticisms or defended the doctrines, because I haven't the slightest idea what the doctrine is supposed to be. Since I don't know what the doctrine is, I can't defend it. If the doctrine is that humans are different from rocks, then I agree – rocks don't learn how to talk. If it is that humans are different from chickens, well, yes, apparently – chickens don't seem to learn how to talk and humans don't seem to learn how to fly. If that's innatism, then I agree with it. I don't understand what there is to be defended, unless somebody really thinks that rocks do learn to talk. There doesn't seem to be any doctrine to be discussed, therefore I don't understand what the arguments are about.

Interestingly, these arguments are invariably dualistic, that is, they only refer to cognitive growth and not to (say) embryological development. Nobody suggests that the transition from an embryo to a chicken is determined by the character of the nutritional environment – like, if you change the nutrition of the chicken embryo's environment, you'd end up with a dog or something. Nobody claims that. And the same is true of everything from conception to birth.

Well, what about after birth? Organisms change after birth too. Nobody, I think, believes that children learn to get bigger. I've never heard anybody suggest

that children are taught puberty – like, they get instructions from their parents that it's time for puberty, or they notice that other kids around their age are doing it so they're subjected to peer pressure and they do it. Nobody has ever suggested that all the way through up to development, except for cognitive development.

Now, it's not that anyone knows anything about these topics. Nobody has the slightest idea what makes people undergo puberty at a certain age, and nobody knows really what makes a chicken embryo turn into a chicken. Some very strange things happen that are hard to explain. Remember that all the cells of the body have the same instructions, but they do different things in different positions. Somehow a particular cell knows at a certain point that it must become a bone instead of a piece of the eye, and to try to figure out how the cell knows that is extremely hard. It's not that anybody knows the answers, it's just that everybody assumes that it is all determined by some inner program, and the outer environment can have at most very marginal influence (can maybe accelerate or retard the development, but can't really change it in any significant fashion), all of this is just taken for granted, even with the absence of any knowledge. And that makes perfect sense because it's obvious from the qualitative character of the problem. You see in all these cases that the organism reaches a very complex, highly articulated state of development, and it does it on the basis of extremely limited instructions from the outside. Therefore, if you're even semi-reasonable, you would assume that it is all inner-directed.

But somehow in the case of cognitive growth, you're not allowed to be reasonable, although the same thing is true here. In fact, in the case of cognitive growth we even know more about it. A good deal is known about the innate structures (more than in many of these other cases), and the qualitative situation is quite the same – you can show quite trivially that people are capable of interpreting and understanding and freely using (without any consciousness of strangeness) all kinds of complex constructions that they've never heard before and so on and so forth. That's trivial to demonstrate. And a fair amount is known about the initial structure that makes it possible (about the uniformities, and so on). Nevertheless, it's considered highly controversial to be reasonable in this domain, and therefore you have debates about innatism, which means taking the point of view that you take towards a chicken embryo. That's very strange.

This is comprehensible on traditional grounds – you can trace it back to traditional religion, soul, and all sorts of other things. But whatever the anthropological explanations (or the explanations as a piece of cultural history), it is the view that seems to have no redeeming features. Nevertheless, it is very strongly maintained. And, again, the most explicit and influential advocate (as in most of

these areas) is Quine. He has put the point across many times over the years. Let me take Quine's most recent version, which appears in a book called *Pursuit of Truth* –

"In psychology, one may or may not be a behaviourist, but in linguistics one has no choice. Each of us learns his language by observing other people's verbal behaviour and having his own faltering verbal behaviour observed and reinforced or corrected by others. We depend strictly on overt behaviour in observable situations."

"There is nothing in linguistic meaning beyond what is to be gleaned from overt behaviour in observable circumstances."

Remember that 'meaning' here just means language generally. We could rephrase Quine as saying that there is nothing in language beyond what is to be gleaned from overt behaviour in observable situations. Accordingly, he says that the behaviourist approach is mandatory.

That's an argument repeated over and over for years, and this is the most recent formulation of it, and it has been very influential. Let's try the argument below the neck. Let's go back to the chicken. I'm going to give an exact paraphrase –

In the passage from the chicken embryo to the mature state, the embryo depends strictly on nutrition from outside. There is nothing in the structure of the mature organism beyond what is to be gleaned from nutritional inputs. Therefore, accordingly, the nutritional approach is mandatory. Biologists must be nutritionists.

Notice that it's the same argument – all I have done is to replace 'observable situations' by 'nutritional input', but the argument is the same and the conclusion, therefore, is that embryologists should abandon this complicated inquiry into how a cell decides to become a bone in one place and an eye somewhere else, and should do something much simpler – just take a look at the nutritional inputs of the embryo (which is going to be really easy because they are the same for a chicken and a frog and everyone else). Well, there must be something wrong with that argument. If I presented that argument to biologists, they wouldn't even bother laughing. Why then is it any more sensible in the study of language than in the study of embryology? That's the question that ought to be asked.

Notice that it's not that the argument is wrong. The argument is entirely correct. It's perfectly true that we depend completely on observable behaviour in observable situations, and it's equally true that the embryo depends on nutritional

inputs only. But the question of course is – what does the word 'depend' mean? How do you spell out the concept of dependence? When you start to spell it out in embryology, you get these complex theories about how cells depending on the gradient of certain chemical concentrations know that they're supposed to produce a specific protein, and after all kinds of complicated stuff you end up with a chicken. Spelling all that out is the meaning of 'depend'. The same problem of spelling out the notion of dependence is also going to arise in language.

Well, there is a theory that spells it out. That's the theory of UG, which says that there's some sort of device that is the internal structure of the language faculty, and it has all these internal parts and properties, and when it gets observable behaviour (what is called data), it spins its wheels in the way linguists try to describe, and it ends up with an I-language – this is the dependence. In embryology, you're allowed to do this. In linguistics, you're not.

It's interesting that in this case, Quine distinguishes psychology from linguistics. In his earlier writings, psychologists had to be behaviourists too –which is connected with the move that distinguishes psychological evidence from linguistic evidence, and psychological evidence has this miraculous character that it enables you to do what scientists do, but linguistic evidence somehow doesn't. I think that this is just pushing the dualism deeper. When you're looking at things like (say) vision, you're not really getting close to the soul of it, therefore you can allow the sciences there. But when you get to language, you're right at the core of things, and here we have to make sure the barriers are very high, and nothing rational is to be allowed.

I don't want to suggest that Quine is religious or anything. In fact, he's a total atheist. He thinks all of this is nonsense and thinks the only thing that exists in the world are elementary particles. Nevertheless, when he deals with these questions, it sort of spells itself out as something very similar to the traditional church. That's an interesting phenomenon about our intellectual culture. The most naturalistic hard-headed scientific people, as they begin to think about things that are close to what traditionally was the "seat of the soul", they veer away from rationality. It's like a particle being repelled by a particle of the same charge. That's a very common feature of our intellectual culture, and it's quite remarkable to see how it shows up even among the most hard-headed scientific people. The phenomenon is definitely worth thinking about for those who are interested in our intellectual culture.

Now, the point is, without properties of the internal states, nothing is going to happen when you're given the nutritional inputs or the observable behaviour. There

must be some structure of the organism, otherwise it's not going to do anything with the inputs. If you give the nutritional inputs to a rock, it's not going to become a chicken, and it won't pick up an I-language, because that happens because there's something about the internal structure of the embryo and the internal structure of the brain.

Quine of course doesn't deny all this. In fact, he even insists on it, and he even tells us what the internal structure must be. He definitely rejects the theory that I described (the one that's guilty of innatism), but he has his own theory about what's in the head. He tells you what capacities the child is allowed to have, and the capacities are those that are spelled out in what is called his 'radical translation paradigm', which is a kind of thought experiment, and if you look at it, the child is allowed to identify objects. The child is somehow granted all of phonology, so it gets words in phonological representations. It's even granted morphology, so the sentences are broken up into words. After it gets expressions like words, the child is allowed to ask a question to see whether people assent or dissent. When the child is shown a certain stimulus and given a certain word, it is allowed to ask the teacher whether this stimulus is a so-and-so or it isn't a so-and-so. The case that Quine discusses, famously, is when you see a rabbit and hear the word 'gavagai', the child is allowed to say, 'Yes, that stimulus is gavagai' or 'No, it isn't a gavagai'. Quine also allows elementary induction to the child. It's not really specified, but if you look more closely, the child is also allowed a quality space – a set of properties like colour, loudness, etc – and you're also allowed a distance measure, so you're allowed to say 'A is closer in colour to B than to C'. Actually, I cheated a little when I said the child sees a rabbit, because of course all it sees is some visual presentation. And in fact, the core of Quine's theory is that there's no way for the child to know whether that visual stimulation is a rabbit, a tail of a rabbit, a rabbit against the background of a bush, a stage in the growth of the rabbit, or whatever. The whole story of radical translation is that anything that's consistent with the stimulation is given by the assent relation. That's the structure of the organism.

So, of course, Quine has a theory of the structure of the organism. He is not saying that it is a blank slate because he recognizes the absurdity in that. But notice that we have gone from one absurdity to another. This is a theory of an organism, and it's an empirical theory (so it must be treated like any other empirical theory), and the theory is so ridiculous that there's no reason to look at it. We know for certain that nothing like that is the structure of the organism. There isn't any reason whatsoever to take that theory seriously. Insofar as we know anything about organisms, they're not anything like that. Therefore, the theory is not worth looking at. The theory is presented purely *a priori*. No evidence is given for it. There's no

indication of how you might proceed to find the evidence for it. It's just stipulated – 'This is the theory and this is what organisms are allowed to have above the neck.' Even if by some miracle it turned out to be true, it would be a totally irrational step. In other fields, you're not allowed to just stipulate *a priori* what the structure of some system is. If someone came along and stipulated the structure of the chicken embryo and by some miracle it turned out to be right, people wouldn't pay any attention to it. If you propose a theory, you must give some arguments, some evidence. Therefore, this is just the purest form of irrationality.

It's true that the theory says that there is a structure to the organism, but there isn't the slightest reason to believe that that structure has anything to do with the organism, and if you try to apply that structure, you get into all sorts of crazy conclusions. You find that you cannot possibly explain what people in fact know. What people in fact know isn't accounted for at all by this theory, and we have overwhelming evidence about what they know. For example, in the case of 'rabbit' or anything else, we know perfectly well that children immediately (actually, on just one presentation) take the word (they don't ask assent-dissent questions) and they instantly know, virtually without error or without later correction, that what was meant was a particular animal and not an animal along with a twig next to it or something. We can accumulate evidence as strong as we like that that happens.

Furthermore, if that was not what happened, nobody would even understand Quine's discussion. If half the people in the world had thought that 'rabbit' means some rabbit-stage and the other half thought that 'rabbit' means a rabbit, how would the first half even understand his example? The fact is that everybody understands his example because we all made the same move. We all picked out the animal as soon as we heard this weird and irrelevant word of Quine.

Furthermore, a fair amount is known about how this happens. About what kinds of entities are picked out (even pre-linguistically) by the child. About what kinds of object constancy there are. About how pre-linguistic children identify stimuli as 3-dimensional and persistent – remaining after they disappear behind the screen and coming out at the right place (they're surprised if it comes out at the wrong place). There are all kinds (not knowledge but) understanding of how children conceive of the world, even before they start learning language. And when they start learning a language, it relates to the way they've already conceived of the world, and as we go on, we find more and more respects in which nothing even happens.

So, we have a stipulation which is completely false, easily refutable, and not worth looking at in the first place because it's totally arbitrary. So why does anybody

pay attention to it? Why is there an influential tradition, including most of serious analytic philosophy till today, that looks at the consequences of the radical translation paradigm?

Now, this paradigm yields a property that Quine calls 'indeterminacy of translation'. The problem is that if you're only allowed this much structure, then it is true that the child cannot tell whether the stimulus was meant to be a rabbit or a stage of a rabbit or whatever, therefore the answer is indeterminate. And Quine therefore concludes that there's no truth to the matter – that the question of whether 'gavagai' means the actual animal or the actual animal with a twig behind it is not a question that has a truth value. It's indeterminate, and indeterminacy is to be distinguished from underdetermination, crucially.

Like everybody, Quine recognizes that everything is underdetermined in the sciences. If you draw a conclusion in the sciences, like, say, in relativity theory, it's of course underdetermined by the evidence, because there were infinitely many possible alternative theories that were consistent with all the evidence around. That's the state of a rational inquiry. The state of rational inquiry is that it's always underdetermined. That's what it means for something to be empirical inquiry. If it's not underdetermined, it's mathematics, not empirical inquiry. The only place where inquiry is not underdetermined is where you stipulate in advance what the answer is, and that's then mathematics. If it's an effort to find out about the world, everything is underdetermined.

But this translation indeterminacy problem is supposed to be beyond underdetermination – something really lethal, something which crucially affects language, meaning, reference, and so on, and then all sorts of conclusions are drawn about language, thinking, people, etc., and we're simply told that we're not allowed to consider an alternative theory of the initial state, namely, the one that works and has scientific evidence supporting it. We're allowed to take this one, which is *a priori* and instantly refuted. That's extremely strange. And we will see how stranger it is if we look more closely into the radical translation paradigm.

Incidentally, the way this is usually described is in terms of the linguist and not the child, and remember, the principle is that the linguist has to follow the path of the child. The child depends only on observable inputs, therefore the linguist has to depend only on observable input. That's kind of strange to start with. The embryologist doesn't have to follow the path of a chicken. If an embryologist is studying a chicken, he doesn't just take in nutritional inputs (like starts drinking the milk or something). So, the embryologist is allowed to do all sorts of other things,

but the linguist isn't. The linguist must follow the path of the child – a hopelessly irrational idea if you think about it, but one that is never questioned in the field. If you look at the literature, this has been around for 40 years and nobody has questioned it internally. There are questions on the periphery of it, but nobody pays any attention to that.

Professional disciplines are very well insulated. That's one of the differences between the sciences and the humanities – a striking difference – that in the sciences, you just can't be irrational, because if you are, then you're refuted at once and nobody cares about you anymore. But in the humanities, you can be as irrational as you like because the disciplines are insulated. They don't really have strong empirical evidence bearing on it. You can go on with this craziness forever. And it happens all the time in the humanities. It's not that the scientists have better genes, it's just that nature is tough and it doesn't let you get away with nonsense. If you try faking an experiment or making up *a priori* theories, people will refute it and then you're finished. Furthermore, in the sciences, you can't keep doing the same thing over and over for the rest of your life. Your students will start asking questions, and suggest something new, and pretty soon they'll be off learning new things and you'll be out of business. In the humanities, that's not true. These are some really big differences.

Actually, it's very striking for me personally. Where I live, there are two major universities within a mile. Harvard is humanities-based. It has the sciences, but the spirit of the place is determined by the humanities – the oligarchy that runs it is from the humanities. The other is MIT, a science-based university. It has some humanities, but they are kind of at the periphery. The difference of intellectual character between the two universities is just astonishing. I've been there for 40 years going up and back between them and it's two totally different worlds, in every area. That's why linguistics is at MIT and not at Harvard. And if you look around the world, linguistics is mostly developed in places where there are no strong humanistic traditions, and it hasn't been developed in major universities that have humanities programs – that's true in the U.S., in Europe, in Japan – and it reflects the very obvious. In the humanities, disciplines are insulated and they don't want to change. They don't want to hear new stuff. They don't want students to come along with challenging questions. They want to keep repeating the same line forever. There's a lot of personal power involved that you can get away with. You're never going to be refuted. You can tell lies about the French revolution forever and you'll never be refuted. And the effect of all this is that when new things come along, they're welcome in the sciences, but in the humanities, it is frightening and therefore they're

pushed aside. I don't want to say this as a hundred percent, but the tendencies are very striking.

Well, let's look at the radical translation experiment and agree to what is completely irrational, namely, that the linguist must follow the path of the child. We have the linguist coming in to study the community from the outside. Here is this 'jungle community', as Quine calls it, and the linguist comes from the outside and is trying to figure out what language they speak. If the linguist were a scientist, what he would do is just what he would do if he was trying to figure out what birds there are in this jungle. The linguist would bring along as part of his package everything that he or she knows about language – and you know all sorts of things about language from the study of other languages, from psychology, etc. We have a huge amount of understanding about language and any linguist who wants to study it would bring along that, just as any ornithologist studying birds would bring it. The linguist would see right away that these things over here are people, just like other people, meaning they're not genetically designed to pick this language rather than another one. Therefore, they fall within the human species. Therefore, whatever is true of Chinese, Japanese, Catalan, Swahili, etc., is going to be true of these people. But the radical translation linguist is not allowed to do this – and that's part of the question of using Wong's information to draw conclusions about Peter. You're only allowed to look at Peter's behaviour.

And notice that that's true of the child too – when the child is trying to pick up this language, it doesn't bring along information about Swahili, but remember that the child has all that in-built and so it doesn't have to bring it along. The child has all that information stored in his brain, and the linguist is trying to find out what the child has stored in his brain. The child doesn't have to bother finding it out – it already knows it. That's why it's really wrong to say that the child doesn't bring information about Swahili, because it does – the child has information about UG. The chicken embryo brings along information about every mature chicken, because it has it stored in its genes, and the same is true of the child.

Well, let's look at the linguist pursuing this curious path – who is not allowed to use anything that's learned about language, not allowed to use anything that might be learned about (say) cellular biology or the structure of the brain or whatever, crucially not allowed to have what the child has, namely, UG stored in its head. The linguist has UG as something he can use, but he doesn't have it as something he can study. Just as the embryologist has genetic instructions, but those don't help him in studying the chicken, and, in fact, the embryologist's genetic instructions are not unlike those of the chicken. I mean, the human arm is like the chicken's arm – the

bone structure is roughly the same, and so on. So, the human embryologist has stored in his genes the answer to the question of how the chicken did what it did, but that doesn't help, because the embryologist is approaching the chicken with his science forming capacities, not with his gene instructions. The embryologist is trying to find out what the gene instructions of the chicken embryo are. The same is true of the linguist. The linguist has UG in his genes, but that doesn't help him in studying UG, for the same reasons.

Now we have the linguist curiously handicapped. He is not allowed to know about what is learned about language. He is not allowed to use information about the initial state that is discovered or stored or whatever. He is only restricted to observable behaviour in observable situations, but, of course, lacking what the child has, namely, the internal structure that made it possible to get something out of those observables. Well, what is the linguist supposed to do? The linguist is supposed to see this stimulus and hear the word 'gavagai', and he is then allowed to ask the informants if that thing is a gavagai or not, and then he gets a yes or no answer, and then linguist is allowed to do simple induction, and you end up with indeterminacy. That's the same course.

Notice what's been presupposed in all this. The linguist was allowed to hear the word 'gavagai', what does that mean? – all the linguist heard was some noise and saw some visual stimulus. How does the linguist get from the noise to 'gavagai'? As any phonologist knows, that's a tough job. Getting from the noise to 'gavagai' involves a lot of theoretical assumptions. Firstly, 'gavagai' is quite an abstract phonological representation. You have to know, for example, if it is the same noise as 'kepeke' that maybe comes along next. Well, that depends on your phonological theory. For example, this could be a language that neutralizes voicing, neutralizes front and back vowels, and has what amounts to the English vowel shift as an option, in which case we would have a repetition of the same word. Those are all sort of intelligible phonological processes.

Well, somehow Quine's linguist has gotten over that stage. He is allowed all these kinds of scientific inquiry that tell you that these are different words (or maybe the same words). All of that's presupposed in setting up the radical translation experiment where you're granted that you hear 'gavagai'. There's an awful lot of linguistic theory granted to the linguist as more equipment, and you get all the same indeterminacy problems there, of course. Hearing the signal and concluding that it was 'gavagai' and not some other thing is exactly as indeterminate as concluding that it is rabbit and not rabbit-stage. That's putting it unfairly, in fact. We have a lot more evidence to support that the child is going to pick rabbit than we do to support

that it's going to pick 'gavagai' by hearing the signal. In fact, the child might have picked different things in the latter case, which is permitted by phonological theory. However, nothing is permitted by the theory of conceptual development, where we have very strong evidence that everyone picks out rabbit and not rabbit-stage – it's not that there are some communities that pick out rabbit and some pick out rabbit-stage. In fact, all that was picked out before the child ever heard any language. In fact, for all we know, it may be built into the child to start with, before any experience.

Looking at the question naturalistically, we've granted exactly the wrong talents to the linguist. We have not granted the linguist the ability to solve the conceptual indeterminacy, although it is solved in the world, and we have granted the linguist the ability to solve the phonological indeterminacy, although it really is indeterminate (though not totally indeterminate –there is a theory which will ultimately solve it, but on presentation of a signal, you wouldn't know whether it falls in one range or a different range). In the conceptual problem, there's no underdetermination at all, as far as we know.

This is very curious. Why is the linguist allowed to do that, and why is he automatically granted all the theory that lets him get as far as 'gavagai'? There's no rational answer, but there's an answer that comes out of intellectual history. 'Gavagai' has to do with sounds of the perceptual articulatory system, and that system is far enough away from the soul to be rational about it, but the concept rabbit is somehow closer to the soul, and we have that same problem of the particle being repelled as it gets closer to the question – as you get closer to the question, you have to veer off into more and more irrationality. I don't think there's any other explanation for this. If you think this explanation is wrong, try to work out another explanation for why everyone's happy with saying there's no underdetermination in the case where there really is, viz., the phonology, and with saying there is underdetermination in the case where there really isn't, viz., the concepts. Why is everybody happy with that, and why is it the basis for huge theories about humans and so on and so forth?

Well, there is a reasonable theory that tries to deal with all these questions (both the 'gavagai'-type and the concept-type questions), and does so with reasonable success. We're told by the anti-innatists that that theory must be abandoned, not because it's false (they don't provide any evidence against it), but it's just that you're not allowed to do it. The philosopher stipulates that you are not allowed to do it. The philosopher stipulates that you have to pick this theory,

although we know it to be false, and totally arbitrary, and the assumptions it makes are completely irrational. That's where that debate lies.

Now, perhaps, it is very likely that what seems to be the right theory ought to be abandoned. The chances that we picked the right theory are very slight, so it will probably be abandoned. Ten years from now, somebody will have a better theory, probably. But to show that it ought to be abandoned in terms of something unknown is not sufficient for the linguist to abandon naturalistic inquiry. It's not enough to say that you have to abandon the methods of the sciences and you have to accept arbitrary stipulations. That's not a good enough argument. In fact, it's no argument. And nothing comparable would ever be considered in the study of other aspects of growth (like the chicken embryo), or even in phonetics, for that matter. If somebody were to come along with similar stipulations in the sound side of language, even the philosophers would laugh. But when we get to the conceptual or the structural side, it's taken very seriously. Again, that seems to be a radical form of anti-naturalistic dualism. To put in the strongest possible terms, nothing can be resurrected from this highly influential picture. It's a paradigm example of what ought to be avoided in philosophy or the sciences or anything else.

We're now dealing with substantive issues, viz., language growth, and I'm trying to argue that the irrationality pervades the substantive issues just as much as it pervades the so-called philosophical issues about attribution of knowledge, etc.

Let's take another kind of substantive issue – the question of reductionism. Natural inquiry leads to all kinds of hypotheses about the brain (of the kind that I roughly sketched out). For example, it leads to phonological theories that tell you what the possible regions are within which that signal can be placed, and how you could proceed to decide what's the right region. It leads to the binding theory, for example, which deals with referential dependency, phrase structure, etc. And it does so by attributing to the brain certain states, certain properties, certain architecture, and so on – well-understood properties like being a generative procedure (which can be spelled out quite carefully). However, this is regarded as highly controversial, and maybe absurd. The reason is that nobody knows how to relate these states and properties to other kinds of descriptions of the brain (say, description of the brain in terms of cells), which is largely true.

Not entirely true, incidentally. A couple of examples have already come up to indicate where it's not true. There are some things about localization that are known that are not trivial. For example, localization of the syntactic and semantic processing for sign language appears to be in the same place as where it is for spoken

language, which is a very remarkable fact because it's in the left (language) hemisphere, whereas visual processing typically happens in the right hemisphere. So, there's something deep about syntactic and semantic processes which is localized in the left hemisphere no matter what the modality is. There are some results by Caramazza on the differences between inflectional and derivational morphology associated with particular kinds of localized brain injury.

There are some other striking examples. Let's have a look at one. This one has to do with some studies recently published on what are called ERPs (event-related potentials). For our purposes, they're just some measure of the electrical activity of the brain. The brain is always producing tons of electrical activity, and it turns out, remarkably, that you can find quite distinctive patterns associated with particular properties of thought and language. For example, it is known for some time that for semantically deviant sentences like "Colourless green ideas sleep furiously" or "I ate the house", when people hear those sentences, the brain produces a specific characteristic electrical pattern which is like a mark of some semantic confusion that took place. Notice that this fact alone would suffice to suggest that the semantic indeterminacy theory is way off the mark, because our notion of semantic coherence is so tight that deviation from it even sets off identifiable electrical activity in the brain. Some more recent work has gone on to find distinctive ERP patterns for several categories of linguistic expressions, which is quite intriguing. It turns out that non-deviant sentences like "John took a walk in the garden" is a category. Semantically deviant sentences like "John took a [crazy word] in the garden" is a category. Phrase structure deviances where you break phrase structure rules is a category. Extraction by relation, of which there are two types of categories – one is the extraction from specific noun phrases. It's well known that "Who did you see a picture of" is an okay sentence, but "Who did you see that picture of" is not good anymore. Something goes wrong here, and the reason has to do with the fact that 'that picture of' is specific in reference – it picks out something particular – whereas 'a picture of' is non-specific. That's a very subtle point, because if you take indefinite noun phrases and give them a specific interpretation, you get the same violation. That's one kind of deviance. Then there's another kind of deviance that's called 'subjacency', and it basically means moving something further than it ought to move. Sentences like "Who did you wonder whether John met" sound worse to people than "Who did you think that John met". It's a well-known difference.

It turns out that these five categories of expressions – non-deviant, semantically deviant, violations of PS rules, violations of extraction from specific NPs, and subjacency violations yield distinctive patterns of electrical activity in the

brain. Each category yields its own pattern. That's quite a surprising result in many ways. These are results that linguists can use to sharpen up their own experiments, because a lot of these judgments are pretty subtle and not terribly reliable.

Remember that our question was of reductionism, which is really the wrong way to put it. The question is of unification. We have a characterization of the brain in terms of computational-representational systems which leads to certain results, explanations, understanding, etc. We have another account of the brain in terms of electrical activity. And here we're beginning to find relations between the two. That's always exciting. But think through what these relations mean. It's very much like the case of click displacement.

These five categories of expressions are well-established within the computational-representational theories. They have a theoretical home. The categories are picked out because of the properties of the theories. There's evidence about them from lots of different languages. In other words, they're embedded in theoretical understanding, therefore they're significant, and all kinds of evidence is bearing on it.

The ERP results are just curiosities. It happened that the people studying the brain included linguists (like Andy Barss) who knew what to look for and they found the numbers, but apart from the correlation, the numbers are just curiosities. There is no theory of the electrical activity of the brain that tells you why these numbers should have any significance. You get all kinds of numbers. These happen to be significant because they correlate with better understood computational-representational theories. So, you have a kind of unification, but it is unification from the weak theory to the strong theory, and now we know something about the electrical activity of the brain, namely, that it seems to correlate with things that really exist, viz., these categories.

I stress this because the standard dualist reaction to this information is going to be just the opposite. The standard dualist reaction is going to be, 'Okay, we've now given some support to the linguist's categories because of the ERP results.' But that's complete nonsense. The ERP results are totally meaningless apart from the fact that they correlate with the linguist's results. They're just curious numbers. Nevertheless, it's a step towards unification, and an interesting one.

[A question is asked by an audience member =] Suppose you have two theories, and one of them correlated with the ERP results and the other didn't. Would that be evidence for choosing between two computational theories?

That's a matter of scientific judgment. You can't give a simple yes or no answer. My scientific judgment is that at this stage the answer is no, because the ERP results are too isolated. They're just funny numbers. It could very well be that if you take that other theory, it will have ten different categories and we'll find numbers correlating with that. Until you come along with a substantive theory of electrical activity of the brain in which these numbers are embedded, the numbers are just curiosities.

It's very much like the click experiments. Fodor interpreted the click experiments to mean that people displace a perceived click to the boundary of the phrase because he knew what the right answer was. If the click was heard in the middle of the phrase, he would've developed a psychological theory which said you displace the click in the middle of the phrase. So, there's no reason to say that the click got displaced to the edge of the phrase, except if you did that you got the right result, and Fodor knew the right result from the linguistic evidence. Now, some day the click experiments might be sharp enough to give you some evidence about something. I think that they probably are – that the correlation to the right result in well-established cases is firm enough so that we can begin to trust the results in the non-obvious cases. In fact, there's some inquiry into that.

Here's a case that's not obvious. If you take sentences like "John expected Bill to be intelligent", there's a real question of where the phrase boundaries are. Semantically speaking, the phrase boundaries are: "John expected | Bill to be intelligent", but the point is that the word 'Bill' acts like it's the object of 'expected' (the upper verb) in many respects. So, it's as if the phrase boundaries are: "John expected Bill | to be intelligent".

That's the kind of case where you might turn to the click experiments to see what they say. Since we have evidence that they give the right answer in other cases, we might be willing to trust their answer in this case. And Bever and Fodor went right on to look at that case because it's the obvious outstanding case. Unfortunately, the results weren't good enough to tell you much, but it's the right kind of thing to do. With ERPs, the same story will be correct. If the ERP results get solid enough, then any linguist will want to take cases where it's not obvious whether you have a subjacency effect or not and see whether you're getting the special result.

Well, that's the stage towards unification. We're relating electrical activity of the brain to the so far better-established theories of computational structure. Despite a certain degree of progress in unification, the fact of the matter is that there are just huge gaps between the brain sciences and the ones that look at the soft stuff in the

brain. And subjects like linguistics (which really are brain sciences, in my opinion, but are looking at it more abstractly, the way matter was looked at by chemists of the 19th century) have huge gaps.

For example, nobody has the slightest idea how a messy object like the brain can create something with the highly refined, extremely sharp, and surprisingly elegant computational properties that seem to be uniformly discovered in natural language. It looks like a very strange thing to come out of a very messy thing like the brain. So strange that a lot of brain scientists just conclude that the linguistics has to be wrong. Gerald Edelman is a good example of a well-known and important brain scientist who simply concludes from this disparity that whatever the evidence is, it has to be wrong – that the brain couldn't do anything like that. I don't think that's a scientifically justified move, because much too little is known about the brain, but there's no doubt that the gap is great. And when you turn to things like the creative aspect of language use, the gap is almost infinite. Nobody has any idea how any mechanism that we understand (quantum-theoretic or whatever), could have properties like ordinary human freedom, or consciousness. There the gaps just seem huge, and nobody knows how to even begin to bridge them.

The existence of those gaps could lead to the conclusion that something's gone wrong, that something's amiss. And if so, we ask where is something amiss? Well, we don't know. In naturalistic inquiry, you hold onto the better-established fields and you start looking for changes in the less well-established ones. That could be the wrong move, we're just making guesses after all, but it's the rational move. In this case, the better-established fields by scientific criteria are the computational-representational ones. The neurological ones are very poorly established – you can find a lot of things about neurons, but you have no idea that they have anything to do with this stuff. You just guess that maybe they do, because they're there. In the case of electrical activity, there isn't much of a well-established theory. Therefore, to the extent that something is amiss, a naturalistic inquiry would tentatively hold on to the quite well-established computational-representational theories and see if there's something radically wrong about the way we're looking at the brain. That would be the rational move. They may be wrong, but, of course, rational moves aren't necessarily right.

Let's look at some analogies to make this look less strange than it may sound. Analogies are never perfect but they can be helpful. One is the classic moment of the beginning of modern science. You had well-established Kepler's laws. You had the mechanical philosophy, which was regarded as self-evident. There was a gap between them, as Newton showed. And what was abandoned was the mechanical

philosophy, which may have been self-evident but was just wrong. That's the step that science took.

Take chemistry and physics of a hundred years ago. You had quite good and well-established abstract theories of matter in terms of things like valence, periodic table, differences between states of matter, etc. You had physics which was taken to be kind of self-evident, namely, particles and fields. There was a huge gap between them, and then it turned out that the physics was wrong, which had to be radically revised in order to be able to incorporate things like the difference between solids and liquids, colour, the chemical bond, valence, etc.

On the other hand, take genetics and biochemistry from a hundred years ago. Genetics was abstract, and there was a huge gap between them. In this case, it turned out the other way round. It turned out by around 1950 that you could account for a large part of known biology in terms of known biochemistry. Here you had what's called reduction. In the other two cases you didn't – you had inflation or expansion or some other thing.

If these examples sound too exotic, let's take one that's very much down to earth, very mundane, which has to do with something called 'nematodes.' Nematodes are tiny little worms who have about 800 cells, a 3-day gestation period, and about 300 neurons.

Credit: Science Photo Library (Steve Gschmeissner)

The wiring diagram of their neurons is completely known – you know exactly how the neurons are hooked up. The developmental pattern is completely known – you know how a nematode gets from 1 cell to 800 cells. Still, nobody has the foggiest idea why a nematode does anything that it does, like why does it turn left, etc. You know everything but you understand nothing about a nematode. Nematodes are critical for this field of biology because it's the most complex organism for which everything is known but nothing is understood. All the wiring is understood, the developmental pattern is understood, but you know nothing about what's going on, so it's an interesting problem.

I'll read you some excerpts from a current research paper on nematodes. Don't bother trying to understand it, I don't understand it either, but I just want to give you the flavour. The guy who wrote it is trying to deal with this gap problem. He says that the way to deal with it is to consider nematodes to be abstract computational devices belonging to a special class of asynchronous interacting automata implementing certain algorithms (which he then spells out) with abstractly-viewed computational and control structures organized in terms of abstract constraints and underlying organizing principles, some of them general (meaning universal

biology), most of them unknown. He takes a highly modular approach, with separate models for developmental structure, and various functions that differ in what he calls 'levels of resolution', from phenomenological to molecular models. He dismisses connectionist models as inadequate for this 300-celled organism, because they abstract much too far from physical reality. He suggests, instead, that neurons be treated as cells that interact by way of a wide variety of chemical information substances, including neurotransmitters, neuromodulators, neuropeptides, etc., acting over multiple characteristic distance- and time-scales, in part, not through synaptic junctions at all. All these crazy systems, he says, are metabolically supported (i.e., they have some interpretation in the metabolism) and are realized in the molecules, but nobody has the foggiest idea how. Putting it differently, this description of nematodes and the explanation of their behaviour in these terms is as yet unconnected with other descriptions (in terms of, say, cells), and the description may involve unification, reduction, expansion, modification, etc. That's nematodes.

Notice that in this case, there are some temptations that no one falls into. I'll quote some temptations from the philosophical literature about language but translate it into nematodes. In this case, nobody says that we have to have Platonic models of nematodes to which the nematode stands in some relation. No one demands that the inquiry has to take account of communities of nematodes. No one claims that there's no truth to the matter because there will be infinitely many theories consistent with the results of some arbitrarily selected experiment or with some stipulation about physical reality. Nobody is tempted to construct a common nematode system to whose principles each worm only partially conforms. Nobody is tempted to hold that if an abstract theory of nematode behaviour is given, it adds nothing to insist that some mechanisms must correspond to the theory.

All of these are quotes from Dummett, Quine, Davidson, and so on, in comparable circumstances about languages. I don't see why their arguments have any more force in the case of language than they do in the case of nematodes. In the case of nematodes, nobody is driven into that direction, so why are we driven in that direction in the case of language? Unless an answer to that is given, we seem to be again in a kind of dualist morass.

The fact of the matter is that whether we're studying organic molecules, nematodes, the language faculty, or whatever, we pursue different levels of explanation, try to find an understanding as well as we can, and hope that we're going to be able to integrate these levels. So, we hope that we're going to be able to integrate chemistry and biology, chemistry and physics, asynchronous interacting automata with cellular biology, I-language with something about the brain, etc. We

don't know in advance whether the right answer is going to be reduction, expansion, change of everything, or something nobody has ever thought of till now. The same is true with electrical activity of the brain and I-language.

There's no point discussing in advance what course the unification will take. We'll know it when we do it. We have no idea how it's going to go while doing it. There are plenty of historical examples in all the directions. As unification proceeds, it may lead to empirical assumptions, including ontological posits, that look absolutely outlandish today, just as it has often been the case in the past ever since Newton. If so, then so be it.

In the case of language and other cognitive functions, it's common to try to relieve the fear that something's amiss, and there are several common ways of relieving that fear. There's a slogan around which says 'the mental is the neurophysiological at a higher level.' That's offered as kind of a definition of the mental. So, don't worry about the mental, it's just the neurophysiological described abstractly. It's a standard slogan.

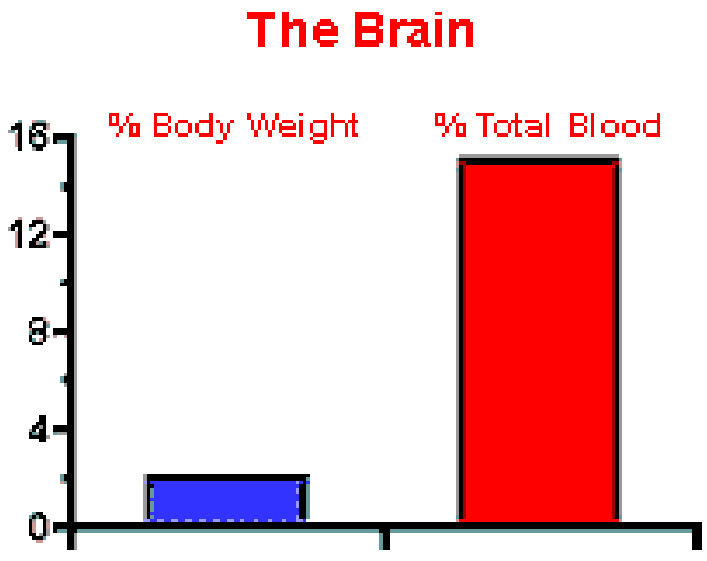
Another approach is that we should eliminate the mental. Just forget it. There's one well-known version of this called 'eliminative materialism' that's associated with the Churchlands, who suggest that people should stop talking about cognitive function altogether and just study neurophysiology. Patricia Churchland wrote a book (an interesting one) that describes neurophysiology for the layperson, and the point is to tell philosophers that this is what we should be studying and forget all the mental stuff like thinking, language, etc. Eliminative materialism is the doctrine that says we should only look at the material world, and not anything else like the mental.

The third approach, which is quite standard, says that we should look at connectionist models. Connectionist models have a different organization – they don't directly do algorithmic processing; they do things in parallel. The systems assume certain forms as a result of continuing interaction with the environment, and they can often do some pretty complicated things. But to say that they don't do algorithmic calculation, which is what is traditionally said, is not a very clear statement, because under some interpretation they are Turing machines, which means they are doing algorithmic calculating. One sort of has a sense of what it means to say they aren't calculating algorithms – they aren't computational-representational in the sense of (say) standard computers. I'll spell this out when we get to the cognitive sciences. But one standard move is to say that the problem with the mental is that you guys are looking at computer models, and you should look at

other kinds like connectionist models, and in that way maybe we can overcome the feeling that something's gone wrong because of this huge gap between the mental (like theories of I-language) and the physical (like theories of cells).

From a naturalistic point of view, all of these are extremely strange moves, and it's worth seeing why. Let's take the first one. We have an account of many things in terms of computational-representational systems, like, we have an account of the phonology of 'gavagai', we have an account of referential dependencies, and all kinds of other things, and they are good accounts with lots of consequences, results, etc. Basically, purely on faith, we assume that there's an account of those things in terms of atoms, cells, etc. Of course, nobody thinks that the operative principles are going to be identifiable at the level of atoms and cells. That's for sure. But people assume that there's some kind of account of all of this in terms of atoms and cells and so on, which is a leap of faith.

With a much greater leap of faith, many people assume that there's going to an account in neurophysiological terms rather than in (say) vascular terms, although the brain is just flowing with blood all over the place. So much blood is in the brain that many biologists who are very skeptical about all of this stuff plausibly argue that maybe Aristotle was right. That maybe the brain is really a thermoregulator. It is there to cool the blood, and then it has these side-functions of language and so on. That's not totally outlandish when you look at the amount of blood in the brain, and there's a lot of other junk in the brain, like glial cells all over the place, and nobody knows what they're for.



Credit: faculty.washington.edu/chudler/vessel.html

Furthermore, neurons are cells – they have all kinds of interactions besides synaptic interactions, as pointed out in the case of nematodes, and the special approach of neurophysiology might just be the wrong one, or the brain might just have other properties that nobody has thought of yet. For example, some of the best contemporary physicists believe that contemporary physics is too misguided to be appropriate for the study of systems like the brain, and that new physics will have to be discovered to be able to deal with such things. We have no idea. The point is that the slogan 'the mental is the neurophysiological at a higher level' intended as the definition of the mental, has the matter backwards. What it should say is that, 'maybe the neurophysiological will turn out to be the mental at a so-called lower level', or maybe not, because maybe it's the wrong place to look. But that would be the naturalistic approach. So, the first approach is wildly dualistic.

What about the second one – eliminative materialism? That's just meaningless. Until the Churchlands tell us what the material world is, we don't know what eliminative materialism is. They seem to understand that cells are in the material world and words aren't, but no scientist can understand that. All we know is that the brain has properties, and some of the properties are that it gives rise to electrical activity like ERPs. Another of its properties is that people who use brains can identify rhyme and anaphoric dependence. There are theories about these properties, some of which are in terms of computational systems. There's no subpart of all of that that's material, at least as far as anybody knows. Until somebody can tell us what the material subpart of all of this, there is no doctrine of eliminative materialism.

Now, we do understand Patricia Churchland when she says linguists and cognitive scientists should stop studying reasoning, thinking, language, etc., and start studying neurophysiology. That makes about as much sense as telling embryologists to stop studying all that stuff and study string theory. Okay, if you're interested in string theory, study string theory. Maybe, ultimately, string theory will account for embryology in some fashion, but to tell embryologists that they should drop their inquiries into what makes a cell decide to become a bone rather than an eye because some physicists have this nice idea about string theory would be totally absurd. To tell linguists that they should study neurophysiology is far more absurd, because it's at least possible that string theory might be the foundation for embryology, and there's some reason to believe that. But there's very little reason to believe that neurophysiology is the foundation for thinking. It could be true, but we have no real evidence for it. So, that's just irrational. Intelligible but irrational. Whereas eliminative materialism isn't even intelligible.

Well, what about the idea that we should move to connectionist models and parallel processing systems? The idea is that – granted that these systems don't work for a 300-celled nervous system with a known wiring diagram, but maybe by some miracle they'll work for 10^{11} neurons, though nobody knows how they're hooked up altogether. Well, maybe, but one would like to hear an argument. It's often argued that we ought to think about the possibility that maybe connectionist models will someday come along that will do what we do in terms of our rule-based models, and we can consider what implications that possibility would have for the existence of rules (and there's a big literature about this). That's very much like saying to the embryologist, "Look, you guys are working out all these really complicated theories about how cells decide what protein to produce on the basis of chemical gradients in the environment and so on, but maybe someday somebody will come along with a completely unstructured theory that doesn't have any of those complications and will explain all the things that you're looking at, so why don't you stop looking at it?" That would be so unreasonable that you couldn't even laugh at it, and it's equally unreasonable to raise the question of what it might imply if some unknown connectionist model (that you can't even dream of) would someday replace rules-based models (which actually work).

Let me stop here. We're right up to the point of getting to cognitive science and its computer models, which is the other standard way of trying to relieve these feelings that something is wrong. That's another wrong move, in my opinion.

Lecture #5

YESTERDAY, I GAVE a few examples of what seem to me to be properly described as non-naturalist, dualist (i.e., methodologically/epistemologically dualist) approaches to substantive issues like issues of fact, issues about the world, etc. Let me stress that the advocates of the position that I'm presenting regard themselves as being paradigm examples of pure, hard-headed, naturalist monism. When I say that it is non-naturalistic epistemological dualism, that is a highly controversial statement, and you therefore ought to take it with a good deal of skepticism since the prevailing view is exactly the opposite, that this is exactly what hard-headed naturalism is. For example, if you read any of the expositions of Quine, or Roger Gibson's books, it is described as the demonstration of what a naturalistic philosophy would be – philosophy that peels away all confusion and residues of the non-scientific path and just pursues questions of language, mind, reality, etc., strictly from a naturalist point of view (it's even called 'naturalized epistemology').

Nevertheless, I'm arguing the opposite, that in fact it is a paradigm example of a new form of dualism much more irrational than the old form – the old form having been incorrect but, nevertheless, naturalistic, and within the framework of the thinking of the natural sciences.

The examples that I gave last time had to do with what's called language learning, but what I prefer to call language growth, i.e., the development of language in the mind. Here the issue is what Quine calls radical translation, which is a certain paradigm that describes a number of different things. The core thing that it describes is the way the child acquires language. The child acquires language by means of the process of radical translation, so Quine assumes.

Anyone who thinks about the problem of language acquisition will recognize that the child is presented with certain data, and something is formed in the child's constitution (in the brain or the mind). Quine isn't very clear about what is formed here, but it seems that what he has in mind is formal language, i.e., a set of well-formed formulas. At least, he says, two different theories about what the child has acquired are indistinguishable if they converge in the set of well-formed formulas.

It appears that what Quine has in mind as the output of the acquisition process is a set of well-formed formulas. Now, that can't really be correct because you can't

have an infinite set in your head. The only thing you can have in your head is a finite characterization of an infinite set. That would be an I-language –a generative procedure that's intensionally characterized in terms of the actual form of the algorithm. Then if you have an I-language, it does generate a class of expressions. Quine doesn't even go this far; he doesn't assume an I-language. I can't even really give an exposition of this because it's incoherent, but if you want to make the view coherent, you would have to assume that there is some kind of generative procedure formed that characterizes the well-formed formulas.

Now, Quine doesn't seem to want to make that move, that is, he regards the question of a generative procedure (what he calls a 'grammar') as being subject to lethal indeterminacy, that is, there's no question of truth or falsity as to which is the right one. So, if there's no question of truth or falsity, then obviously it can't be in the child's head, because if it is, then it's just as real as (say) his arms and legs. Therefore, if the grammar is subject to lethal indeterminacy, it cannot be in the child's head. Therefore, we somehow are going to a class of well-formed formulas without any finite characterization of them, and then we're making guesses about possible finite characterizations.

Now, this is all completely incoherent. I'm repeating the words but they don't mean anything, because you can't have an infinite set in your head. Beyond that, there's the question of whether the formal language even exists. As far as I know, it doesn't exist anyway. There's no evidence that there is such an object. There's no gap in linguistic theory that's waiting to be filled by the invention of such an object. No one has ever made a proposal as to what such an object would be. So, it's a pure mystery, but that's somehow the picture.

There is a coherent picture, but Quine won't accept it. The coherent picture is the naturalistic one that I described before. The child gets the data. There's an I-language – a very specific generative procedure. The I-language is a real object, just as real as the visual cortex, and we may use any evidence at all to try to find out what it is (biological evidence, evidence from other languages, quantum physics, etc). We just follow the methods of science, which are completely opportunistic and look anywhere for evidence.

But we're not allowed to do that under the paradigm of radical translation. The picture here is that a child goes from data to a class of well-formed formulas somehow, without going through the generative procedure that characterizes it. You can't make any sense out of this. Something's missing, and for 40 years I've been asking Quine what's missing, and I never get an answer, and there's nothing in the

literature that tells you what the answer is. That's where it stands. I can't give any clearer a characterization of this. And this characterization is incoherent too, but is something intrinsic to this approach, as far as I can see. If somebody knows how to do it, I'd like to hear it, but I've never heard it and can't find anything in the literature about it.

Somehow the child goes from the data to the class of signals or well-formed formulas or something like that, and Quine recognizes, of course, that unless there's some structure to the child, he won't do anything with the data. Quine then simply stipulates the structure of the child. He doesn't regard it as a matter that requires evidence. That's a curious move. You're talking about something real – the structure of the child. If someone makes a proposal about the visual system saying that it has such-and-such property, they're expected to give evidence, without which nobody will pay any attention to them. In this case, somehow you don't have to give evidence, you just stipulate it. And the stipulations are that there's a quality space that has a kind of a metric in it, meaning there are the elementary properties of things, like, colour, size, shape, loudness, etc., and some measure that tells you how close things are, which allows a degree of generalization. Then there is conditioning. There is the assent-dissent experiment that the child is allowed – the child is allowed one experiment, namely, asking if this is a so-and-so or not. There's elementary induction, meaning simple induction – if a thing has happened a lot of times, it will happen again, that kind of thing. And, tacitly, there is some linguistics, and you don't know exactly what, because it's never stated, but it's at least phonology. And, apparently, also morphology – now we're talking about interpreting texts, that is, you see what's written and try to figure out what the person has in mind – and it seems that if you look at the text, you find presuppositions about morphological structure – inflections, plurals, etc – and they seem to come from somewhere, and the child has them, so, apparently, the child has morphology. The child is somehow allowed to pick up words, and maybe something about sentences too (I'm not sure), but some range of things are given tacitly and we're not told why. That's the radical translation paradigm. The next crucial aspect to it is that the linguist is identified with the child.

Actually, there's a further point which I didn't mention yesterday that I'll talk about now – the person in a communication situation (like if you and I were talking) is also identified with the child. So, all three of them (the child, the linguist, the communicator) require the use of the radical translation paradigm.

With regard to the child, the radical translation paradigm is a theory, or it would be if it was filled out and if the incoherence was removed – if you could

explain what you mean by going to the set of well-formed formulas, and if you could fill in these tacit assumptions, and could spell out the quality space, and so on and so forth, then you'd have a theory of the child, right or wrong. In fact, so obviously wrong that no scientist would even look at it, as is always going to be the case when you simply stipulate a theory. If I were off the top of my head to stipulate a theory of the visual system, the chances that it would have anything to do with reality would be zero. And the chances that this one has anything to do with reality are also zero, and insofar as it's clear, it's just flat wrong in every respect – that's why nobody who works in language acquisition pays any attention to this model. But strikingly, this theory is arbitrary and stipulated, and that itself is a very sharp departure from naturalism.

Now, when I say it's arbitrary, I don't mean that it has no historical sources. And here we get into the matter of hermeneutics or something – trying to figure out what's in the back of people's minds. And in Quine's case, it's not very obscure. What's in the back of Quine's mind is, I suppose, formal arithmetic, or other formal systems like arithmetic.

If you take a course in metamathematics, you'll start off by defining something called a language (although it has none of the properties of a natural language and is just a metaphor, and, in fact, is radically different from any natural language in all respects), and the language is formed by picking some symbols, like, plus, zero, successor, etc., (that's the phonology, if you like) and then you define a mathematical operation called concatenation that allows to string these symbols together, and then you give a criterion which picks out a class of well-formed formulas – you say that among all these strings, there's a certain infinite class which are the well-formed formulas (like *2+3=17* is a well-formed formula, but *(+)=3* isn't), and you give a criterion for that, which is a mathematical operation like (say) defining the set of even numbers. And that set of well-formed formulas is your language, and then you can pick one or another generative procedure to characterize that set, and indeed it is indeterminate –because it doesn't matter which generative procedure you picked.

Within the class of well-formed formulas, you may want to make a further distinction between those that are theorems and those that are not theorems. So, *2+3=17* isn't a theorem, but *2+3=5* is. Then that class of theorems can also be characterized by some procedure, and that procedure is called an axiom system (and as long as it gives the right class of theorems, you can pick it any way you like).

Then you can start asking very interesting questions. For example, you can give an interpretation of this system under which some well-formed formulas are true, and you can ask whether the truths are the same as the theorems, and when you pursue these questions, you reach some quite spectacular mathematical discoveries – the great discoveries of modern mathematics like the Gödel theorems, etc. So, by no means this is a trivial subject. It's a major subject in modern mathematics, with enormous implications.

If that's the model you have in mind, then you can sort of understand what Quine's thinking about. In that model, it's quite true that you specify the language as a class of well-formed formulas, and then you pick the grammar (the generative procedure) any way you feel like, and what the theorems are depends on things like beliefs about the world, etc., and they can be modified as your beliefs are modified – there's no definite set of theorems because that depends on your choice of the axiom system.

Even for systems as simple as arithmetic, it's a remarkable fact that there is no way to give an axiom system that captures our notion of the natural number. Any axiom system that you construct is consistent with what are called non-standard models, i.e., classes of things which don't have the properties of natural numbers.

You can see why people who know all this stuff would be led to theories of indeterminacy for something as complicated as language, but the trouble is that it has nothing to do with natural language. It has no relationship at all, from the very first step. The structure of formal arithmetic is radically unlike the structure of natural language. Furthermore, these are creations of the mind. Natural language isn't a creation of the mind any more than your arms are a creation of the mind – it's just a form that the brain takes under certain conditions, it's not something you create by thinking about it. And it's interesting to know how people create things, and you can ask interesting questions about what the nature of mathematical truth is, but that's a different field. The study of language is like embryology – it's the study of how the brain takes a certain form under certain conditions – and any mathematical analogies are totally irrelevant. People think there's a formal language and a natural language, I suppose, because there is a formal language in arithmetic, but that's because you stipulate it to be so. Just like if you're doing number theory, you could stipulate the class of prime numbers, and it exists because you stipulated it, but you can't stipulate things in the natural world – they're either there or not there.

Therefore, carrying over these analogies and models is very misleading. Remember, what I've just been saying is an interpretation of what I think is in the

back of the minds of people who create these systems. They don't say so, but I'm just guessing as to what's in the back of their minds which could lead to a picture so totally incoherent, arbitrary, stipulative, and anti-naturalist. So, I'm trying to make up a story that would make it intelligible and understandable that smart people who know mathematics could be so misled when they're studying the natural world. Maybe that's the right story or maybe it's the wrong story, but that's a matter for biographers – those are the questions that ought to be raised in intellectual biography if it was ever done by people who aren't just hagiographers (which is rare). Anyhow, that's the picture, whatever the source may be. Now, we have to somehow try to make sense out of the picture, since it doesn't make sense itself.

The picture has various problems. One problem is that the output, the class of well-formed formulas, is a completely crazy output – there's no such thing, as far as we know, and there couldn't be such a thing as an output unless there was an explicit, fixed, intensional characterization of it (an I-language), and since the existence of the I-language is denied, we're in total incoherence. So, firstly, the output of the system is crazy.

The second thing that's problematic about the theory of radical translation is that the internal state is crazy, not because it couldn't be that way. The output couldn't be that way because there couldn't be a possible organism that stores an infinite set without a finite characterization of it, so that's just incoherent. But the internal structure is not incoherent – you could design an organism which has those properties – it's just that there isn't any such organism around, in fact, none remotely resembling it. What's crazy about the internal state is the idea of stipulating an internal state. The internal state of a complex organism is a very hard thing to understand. If you want to understand what makes the chicken embryo turn into a chicken, you have to do a lot of work. If you want to understand what causes a human to grow arms and legs and not wings, you have to do a lot of work. And you're not going to have to do less work than that to determine what makes the human brain turn into a human language. So, the idea that you could stipulate all these things is just outlandish.

The third thing that's crazy is the identification of the linguist with the child, because their situations are totally different. The linguist coming to investigate the 'jungle language', as Quine puts it, is not at all in the epistemic situation of the child. The epistemic situation of the child is that he comes already equipped with the innate knowledge of what a possible language is. The child comes to the language acquisition problem knowing, tacitly, that there's a certain small class of possible answers, from which it picks out an answer – an I-language. That's the epistemic

problem for the child. The epistemic problem for the linguist is totally different – the linguist is trying to find out what is the epistemic state of the child. That's the linguist's problem.

So, on the one hand, the linguist does want to find out what the child ends up with, like, you want to know what's the structure of Catalan, but the more interesting question is what is the structure of language, which is a question about UG, which is simply the question of what the initial state of the child is. So, the task of the linguist is to discover the epistemic state that the child begins with. To insist that the linguist approach that problem by being in the epistemic state of the child doesn't make any sense.

In fact, the linguist will use any data he can find. Why should the linguist be restricted to this data? The linguist will, certainly, use data from other languages, and, in fact, every linguist always does that. If somebody is studying Catalan, and if they have two choices as to how to handle causative constructions, they'll ask questions about how these constructions work in other languages. That makes perfect sense, because however they work, it's being determined by the internal state, and if you have evidence about it from Japanese then that's going to bear on Catalan. If anything were discovered about (say) ERPs, or the theory of cells, the linguists would be happy to use that as evidence. You're certainly not restricted to this database, and you're not in the epistemic situation of the child. A person inquiring into the nature of the epistemic state isn't in that epistemic state. He's using some other capacities of the mind, the capacities that we use when we try to understand (say) quantum physics. When you're trying to understand quantum physics, you don't use UG, which is just irrelevant. You use whatever the relevant aspects of our science forming abilities are, and those are the very same aspects that you'll use in trying to answer this question of what's this epistemic state.

So, the picture has three forms of craziness, if you like. One – the output, which is incoherent. Two – the input, which is stipulative and wrong. Three – the assumption that the linguist is in the epistemic state of the child, which doesn't make any sense at all. So, it seems to me that this entire picture has nothing in it that can be resurrected, and any conclusions that follow from it are valueless.

Now, very far-reaching conclusions are drawn from this, about the theory of meaning, about people, etc., and it's a highly influential paradigm, but, as far as I can see, it's about as remote from naturalism in every respect you can imagine.

Let me finally add the thing that I forgot to mention last time – the identification of the person in the communication situation with a child. The idea is that we have two people talking, Peter and Mary, and the question is how do they proceed to communicate. According to the Quinean view, they essentially proceed by radical translation. So, Peter has this set of qualities and Mary has that set of qualities, and maybe they've already achieved some kind of language (which is not clear since it's not stated, so I don't know what to say about that), and then they try to communicate – Peter hears Mary say 'gavagai' and tries to figure out whether it means rabbit or rabbit-stage. That's the story. So, the person in the communication situation is like the child.

Notice again that this is radically different from the way we study any other kind of communication. For example, there are questions about how cells communicate (they do interact in all kinds of ways), and you never study cellular interaction by stipulating something like a radical translation paradigm and saying that that that's how it works because someone says so. It would be totally idiotic to even consider that.

There are all kinds of interesting work in biology that studies communication between systems, and the way it works is always the same – you try to discover the internal state of the organism, and then you ask two questions about it – 'What is it?' and 'How did it get there?' So, if you're studying a communication situation, you want to know what's the internal state of the entity and how did it arise. That's standard in science. And, in fact, that's the way it is done for cells, ants, chimps, birds and so on. But when we get to humans, we're not allowed to do that anymore. The same dualism suddenly arises. When we get to humans, we're not allowed to ask what are the internal states of Peter and Mary, and how do those internal states affect their interaction, and how do those internal states develop. The questions we would ask about everything else from cells up to humans, we're not allowed to ask here. Rather, we have to assume that the epistemic state of Peter and Mary is what is stipulated by the philosopher.

There's nothing in the history of philosophy as irrational as this. I mean, there are a lot of irrational things in the history of philosophy going back for thousands of years, people have made stipulations about the natural world, but I can't think of anything as irrational as this. If you go back to the earliest period, the stipulations about the natural world were speculations based on current understanding. Democritus and his atom came out of speculation based on current understanding – you saw that if you poured oil on water, it didn't extend forever, and so you think there must be discrete particles or something (which is quite good reasoning). Or

when David Hume speculated that the springs and origins of human understanding, as he called it, involved certain principles of association and so on (which is wrong but not unreasonable). I think you have to look very far to find anything as irrational as this. I think it's unique in the history of thought in the level of irrationality that it's reached, and, strikingly, there is no internal critique of it, and very sophisticated and smart people pursue it as if it makes sense, without raising any serious questions about it. If what I'm saying is correct, it does require explanation.

There's another strange aspect of this – that is the willingness, in the case of 'gavagai' (which is the standard example), to allow the child, the linguist and Peter in the communication situation to pick out the phonology of the word, but they're not allowed to pick out the meaning of the word – rabbit – rather there's indeterminacy about the meaning. That's an asymmetry that requires explanation, because picking out the phonology of the word is not a trivial matter, and, in fact, there really is underdetermination here. Even just writing the phonetic transcription of the signal already brings to bear a rich theoretical apparatus based on universal phonetics. But even if we get beyond that, we know perfectly well that what range of variation falls within what we will ultimately call the word 'gavagai' for a particular language is underdetermined, and certainly underdetermined by the first signal – it requires building up a pattern, and a system, and meeting certain formal conditions, and so on. So, there are some problems there, but somehow we're allowed to get over those, and maybe over the morphological ones too.

But, again, no reason is given as to why we're allowed to do those things but not do what amount to comparable things on the semantic side. On the semantic side, there's supposed to be lethal indeterminacy – there's no truth to the question whether 'gavagai' means rabbit or rabbit-stage, but there is truth to the question whether /gavagai/ or /kepeke/ are the same word. This requires explanation, because they are similar problems, and we probably have better evidence in the case of rabbit and rabbit-stage than in the other case. But that question is not addressed, and we might ask why, and again we have to just try exegesis – and my guess is that there's a tendency as you get closer to the traditional soul to veer off into irrationality, for whatever reason. Therefore, you kind of presuppose the rational assumption in the case of the phonetics, which is far enough away not to bother anybody, but you're not allowed to presuppose the rational steps when you get closer to semantics.

An argument is given, in fact, at this point. The argument that's given is the theory of 'meaning holism'. This is a notion that comes in various strengths, depending on how holistic meaning is. The least strength is (which almost everybody in the field accepts) that the word doesn't have meaning in isolation but only in the

context of the sentence. So, a weak version of meaning holism is that only in the context of a sentence does a word have a meaning. So, you first learn sentences and then kind of abstract away the meaning of the words. A richer theory is that a word has meaning only in the context of a language. A still richer theory is that a word has meaning only in the context of (what's sometimes called) a theory – a total belief system of which language is a part. Quine, crucially, believes that a word has meaning only in the context of a theory – that's the extreme version of meaning holism. And that could be true. After all, that's a substantive proposal. If it is true, that would give an argument for the distinction between sound and meaning, because sound is not supposed to be subject to this kind of holism. But we might ask whether sound could be subjected to this kind of holism.

For example, there's an interesting book by Jerry Fodor and Ernie Lepore called *Holism: A Shopper's Guide*. The book is a critical analysis of meaning holism. So, they're not advocating but just considering various reasons why the idea looks plausible, and showing problems with those reasons. They start the book by saying why meaning holism looks so plausible. In fact, they say that it even looks obvious. Setting forth their own project, they say, 'Look, here's this idea that looks totally obvious, but, nevertheless, we're going to show some problems with it'.

Take the word 'bark' – it obviously means different things if I say "I peeled the bark off the tree" than "The dog barked". Therefore, we have shown that the meaning of the word is determined by the sentence. Or take 'flying planes.' If you put it in the context 'is dangerous', it means one thing, and if you put it in the context 'are dangerous', it means another thing. Therefore, the meaning of the word depends on the sentence, and we've established the weak form of meaning holism.

What about the stronger form? They consider the expression "Empedocles leaped". They say that if you take this sentence in English, it means that the philosopher jumped off the mountain, and if you take it in German, it means that the philosopher loved (*geliebt*) somebody. Therefore, we've shown that the meaning of a word is dependent on the language. There are various points finessed here, like why "Empedocles leaped" is the same sentence in English and German? But let's forget that.

What about the dependence on the theory? Here there are standard arguments given by everybody. They say, take the word 'momentum'. At one time it meant mass times velocity, and at a later stage of physics it was just given a different meaning. In fact, the definition of it in the early stage turns out to be false in the new stage – so that's a radical change in meaning as theory changes. Or take the word

‘atom’. It meant one thing to Democritus, something else to Dalton, something else to Niels Bohr, and something else in quantum theory. So, the meaning of the word changes as time goes on, and therefore, the meaning of the word is dependent on the entire belief system.

The question is, how strong an argument is this? See, once you’ve established meaning holism, then you’ve essentially said that the study of meaning is hopeless, because if the meaning of the word is going to depend on sentences, languages, whole theories, and so on, then just forget it – it’s like saying that in order to find the meaning of a word, you’ll have to know everything. That’s the same as saying that there’s no study, that the topic is finished, and now we have demonstrated that there is a real difference between sound and meaning. And that’s what lies behind all this stuff – the radical translation paradigm is supposed to give you the rationale for the theory of meaning holism.

Let’s try to construct a theory of sound holism and see what it would look like. Firstly, what do people mean when they say the word ‘bark’ has a different meaning in different sentences? What’s the word that has the different meaning? Presumably, by ‘word’ they mean the collection of phonological properties (or maybe the letters or something). Presumably, the output side of the word – the signal side – is what they mean by the word, but that’s a funny notion of word. The word ‘bark’ in English isn’t just a bunch of sounds, it is a collection of a bunch of sounds and a bunch of meanings – there’s a Saussurean connection between the sound and the meaning – and you have no word unless you have that connection. So, a word is what I call a ‘linguistic expression’, which is just a collection of properties – sound properties, meaning properties, structural properties (‘bark’ is either a noun or a verb but not an adjective), etc.

Now, if you extract the meaning properties, then you can run through this argument and get a theory of meaning holism, but just mechanically we could do it the other way around – we could leave the meaning properties and cut out the sound properties. Now it’s harder to talk about it because it’s harder to talk about a word without giving it a sound, but that’s just a hurdle. Then we could say that the non-phonological word ‘bark’ (‘bark’ with all its properties intact but divested of the fact that it begins with a bilabial stop and so on – divested of its sound properties) has a different sound depending on where it’s placed in the sentence (just as we said that ‘bark’ divested of the meaning properties has different meaning in various sentences), which is in fact true – ‘bark’ sounds different if you utter it at the beginning of the sentence than at the end. Therefore, the non-phonological ‘bark’ depends on where it is in the sentence. Similarly, in “Empedocles leaped”, if we take

the entire conceptual structure of the philosopher jumping off the mountain and divorce it from the sound part, we can say that it sounds one way in English and sounds a different way in Catalan, therefore the sound of the sentence depends on the language (and we could go up to theories, if we like).

So, we've got a theory of sound holism which says that the sound of the word at least depends on the entire language. That means it's completely hopeless to study sound. So, we can throw away phonology because in order to study the sound of the word we have to learn a whole language. Obviously, that's absurd, and the argument is no better in the other case. So, we have no argument for meaning holism, because it is crucially based on taking a word to be a collection of properties minus the meaning, and if you do that then you can establish meaning holism. But if you take the properties of a word minus the sound, you can establish sound holism in the same fashion. But that's just a trick. Since we don't take sound holism seriously to conclude that phonology is a hopeless subject because you have to know the whole language or the whole theory before you study the sound of a single word, there's no reason to take meaning holism seriously, at least on those grounds. Again, that doesn't mean that it's false – the fact that an argument for X doesn't work doesn't make X false.

What about the last part that words change meaning as theories change? Well, that's a notion that's extremely hard to even state. When we say that 'atom' meant one thing for Democritus and something else for Niels Bohr, what exactly are we saying? What is that word 'atom' that was shared by Democritus and Niels Bohr? There's no real answer to that – they talked in different languages, they put it in different contexts, etc – it's very mystical notion. This relates to something I mentioned on the first day, about the free and easy talk that people give about words changing meaning or changing sound, etc. That doesn't really mean anything much. I mean, you can get away with it in normal discourse where precision isn't very important. But we're out of normal discourse now, we're trying to give a theoretical account of something, and when you do that, you have to be much more careful about what you're saying. Now you have to give a sense to the concept that words change meaning or sound, and there is no sense to that concept.

Notice that there is an interpretation of all this in I-language terms that makes perfect sense. To say that a word changes meaning over time is to say that if you take successive I-languages and you draw a line through that particular word, you'll find that that element has different positions in the system at different times. It's a bit like saying giraffes evolved longer necks. We can say that informally, but it doesn't mean that there was a thing called the giraffe that had a short neck once and

has a long neck later, of course. What it means is that the distribution of neck lengths was such and such at one period, and it was some other thing in some other period. That's what we're saying when we say giraffes evolved longer necks. As long as we're not confused about the matter, there's nothing wrong with saying giraffes evolved longer necks, but if some philosopher came along and said that there's an entity called giraffe and it used to have short necks and now it has long necks, and a process called 'evolution' took place between them, that philosopher would be going into outer space. He would be entering into total confusion, because these are just shorthand ways of saying that the distribution of traits changed over time. That's what's involved when you say that words change meaning, sound, or something else. If somebody asked whether it is the same giraffe now as it was a couple millennia ago, that wouldn't mean anything. I mean, all we can say is that there were a bunch of organisms then and there are a bunch now and there's a strange complicated historical connection between them and that's the end of the story. There's no question as to whether today's giraffe is some version of early giraffes.

Similarly, to ask whether 'atom' for Bohr is the same as it for Democritus is just not a question – they spoke different I-languages which matched in certain respects, and we'll find certain similarities between them, and maybe those similarities will be enough so that we'll decide to pick a point in one system and say that it is roughly the same as the point in another system, but that's all you can say. It's like giraffes and necks.

So, the entire question is being posed in a way which doesn't make a lot of sense and requires a good deal of clarification, and it's clear how to give a clarification in I-language terms. But once you give the clarification, the problem dissolves, and now there's no longer any question about whether the word changed, because it's not the same word any more than it's the same mountain after an avalanche, which is not a meaningful question – it depends on what your interests are. I mean, it's obviously not the same physical object – if your current concerns as to how to individuate mountains are such that the physical object before and after an avalanche are similar enough, then it's the same mountain for you, if they're not similar enough for your current concerns, then they're not the same. But the question is a question of decision, not fact.

The second point to notice about these examples of theory holism is that the examples that are given are almost always from natural science. So, it's not the word 'fall' that's given. Let's take a sentence that we use in ordinary English even when talking about a relatively technical problem. So, suppose I say "the missile rose from the ground and then fell back to the ground". Notice that the meaning of the words

is not part of natural science. Firstly, there's no such thing in the natural world as 'the ground'. If you ask a geologist where 'the ground' is, he wouldn't know what to say. Suppose I make it even worse and say "the missile rose toward the heavens and fell back to the ground", there certainly isn't anything called 'the heavens.'

So, in "the missile rose to the heavens and fell back to the ground", there's no 'heavens' and there's no 'ground'. Furthermore, 'rose' is wrong because what happened is that the missile pushed the earth down a little bit, and the earth pushed the missile up a little bit, and since the missile has much less mass than the earth, the difference from the initial position is greater for the missile than for the earth. So, they just pushed each other away. And when the missile fell, all that happened was that they pulled each other, and if you really get fancy, what happened is some geodesic through curved spacetime. Nevertheless, we understand all this fine.

Notice that the words of ordinary English didn't change meaning as theories changed. I mean, there was a time when people would've meant that there's a thing up there called 'the heavens' and there's a thing down here called 'the ground', and the missile went from the ground to the heavens and back to the ground. There was a time when that was science. It's not our science now, our theories have changed radically. In the scientific part of our minds, if we're fancy, we say something about geodesics and curved spacetime, and if we're not that fancy, we say that a collection of mass particles and another collection of mass particles pushed each other apart and pulled each other back together, but the language hasn't changed at all – none of the meanings of the words have changed as theory changed, and that's quite typical.

In contrast, there is no longer any term in scientific language that corresponds to what Democritus meant by 'atom' – the language of science has no such term anymore. In fact, it has no such term as what Bohr meant by 'atom', because the theory has changed enough so that that notion isn't around anymore and some other one is around. We could say that that new notion is kind of a descendant of Bohr's 'atom' just as today's giraffe may be a descendant of some earlier giraffe, but that's just a metaphorical shorthand for something much more complicated – something about the comparability of points in two different systems.

The point of examples like these is to illustrate that the argument about meaning holism based on change of theory works for natural science only if we interpret 'change of theory' rationally – that we give up certain forms of theory and adopt new forms of theory, sometimes using the same word, and sometimes with enough similarity between them (and there's no metric to this since it's just a matter

of decision) to say that the new atom is a new version of the old atom, just like we might say that the mountain after the avalanche is the same as before the avalanche.

On the other hand, for natural language, the theory of meaning holism just seems to be flat false. If we look at natural language, we see, in fact, that the terms have retained all their old meanings as theory has radically changed, and it doesn't matter how sophisticated a physicist you are, you'll still say, "the missile rose from the ground to the heavens and fell back to the earth" and mean exactly that, because the language hasn't changed at all even though the theory has changed totally. Insofar as the strongest theory of meaning holism (the one that makes the meaning of the word theory-relative) is intended to be a substantive comment about natural language, it just isn't true. Insofar as it's a comment about scientific language, it may be true but only in the sense in which giraffes evolving longer necks is true.

These issues arise very crucially in connection with contemporary theories of reference – what are called 'externalist' theories of reference – theories that tell you that the reference of a term like 'book' or 'table' is determined by the way the world really is, that reference and meaning aren't things in your head, they're things in the world. Hilary Putnam is one of the originators of these theories, which sort of swept the field in the past 20 years. Putnam, in talking about the theory, gives the example of 'momentum'. He says that momentum is a particular thing they defined in the 18th century, and they were just wrong, because they didn't capture that thing (contemporary physics defines it a different way and they may be right). But he says that there's a thing called momentum in the world and a word 'momentum' that refers to that thing, and people who use the word may have been misreferring all the time.

For Putnam, there's a real substantive question as to what 'atom' meant (say) pre-Dalton, and Putnam says that what it meant is what we mean, it's just that he didn't know. Niels Bohr, when he constructed his model, meant by 'atom' what we mean, even though he misstated its properties. Putnam argues this on the following grounds. We have to assume this, otherwise we would be forced to say that all of Bohr's statements were false, and all his beliefs were false, and it would be as if he was doing astrology. Well, obviously, Bohr wasn't doing astrology. So, unless he meant by 'atom' what we mean by 'atom', we can't even interpret his theory in our terms, because it's not the same words – it's as if he's talking Greek and I'm talking Swahili, and we don't want to say that because obviously he was talking sense. Therefore, we have to conclude that by 'atom' he meant what we mean by 'atom', and some of the things he said about atoms are true and others were false, but that's understandable. Therefore, we have to take this view that says that a word has a fixed

meaning determined by the world – like what a tiger *is* is determined by the nature of the world – and it doesn't matter what concept we have in our minds. We're forced to conclude that in order to account for the notion of intelligibility in science. That's basically Putnam's argument.

Notice that there are a couple of crucial assumptions that ought to be brought out and questioned. One is that scientific language is part of natural language (talk about electrons, momentum, etc., is part of natural language), that is, there is no fundamental difference between concepts like atom and concepts like house. And that's very dubious, as you can see from the "the missile rose to the heavens and fell back to the ground" example, which is part of natural language, and all the considerations that lead us to consider that things are really pushing and pulling each other apart just don't bear on natural language – that's another part of our minds where we construct sciences.

When we construct the sciences, we often use the resources of our natural language, because that's what's around, but we're changing them in all sorts of ways. In the natural sciences, we are constructing symbol systems, but those symbol systems can include, for example, calculus, differential equations, etc., which certainly aren't part of natural language. Furthermore, in that symbol system that we're constructing, we're trying to impose certain properties upon it. For example, one of the things we're trying to impose on this symbol system is the property that there are words, and that these words refer to real things, and among those things are natural-kind things (the things into which nature is really subdivided). In short, we want to construct this symbolic system so that it has realistic properties.

The purpose of the game of science is to construct a symbolic system which will have whatever resources it needs (like differential equations), and will have a relation of reference holding between words and things, and if you discover that the concept that you've proposed doesn't refer to a thing because there is no thing of that kind, then you change the concept. And the natural kinds have to be real kinds. For example, for the ancients the real kinds were earth, air, fire, and water. But those aren't real kinds for us any longer. For contemporary science, there are no terms referring to those that the ancients referred to, because earth, air, and fire aren't a kind. We're consciously aiming in science construction to develop a symbol system with realistic properties which picks out natural kinds.

But natural language has none of these properties. In fact, in natural language, words don't even refer. That's not the way natural language works. It's true that people use words to refer, but that's something different. People use words to lie,

but that doesn't mean the words lie, and it would be a mistake to think that. The idea that words refer is a huge leap, a leap which comes naturally to people who think about the sciences, because in the sciences you're trying to construct words that refer. Also, words in natural language don't have any particular relation to natural kinds.

It is kind of historically interesting to note that in the contemporary discussion of these issues, people talk about 'water' and ask 'What did some 16th century person mean by 'water'?', and the answer is what we mean by it – H2O – and the person didn't know it but they meant H2O. Now, it's interesting that nobody ever picks 'earth', 'air' and 'fire', and asks what they referred to. For the ancients, earth, air, fire, and water were on a par. When you read contemporary philosophical literature, why do you talk about 'water' and not 'earth'? The reason is obvious – there's a rough counterpart in the natural sciences to our intuitive concept of 'water', namely, H2O. There isn't even a rough counterpart to the other terms, so you don't use those. But the point is that it's only a rough counterpart. There's no reason to believe at all that people in the 18th century (or people today) meant H2O by 'water'.

If we're talking as scientists, we probably mean H2O. But if I say "I'll drink this glass of water", I'll drink it, though I know perfectly well that this isn't H2O – it has some H2O in it but an awful lot of other junk with it. Yet we refer to it as 'water' and not (say) 'milk', and that's correct. I use the word 'water' to refer to this glass of water even though I know perfectly well it's not H2O. The reason why in the contemporary literature you get discussions of 'water' and not 'earth' is that there isn't even any rough counterpart to 'earth' in the natural sciences.

The right lesson to draw from all of this is that the symbolic systems constructed by the natural sciences are just unrelated to natural language, just as the formal systems constructed in formal arithmetic are unrelated to the natural languages. Well, 'unrelated' is a strong term, because it may be that they're influenced by them in some respect, but to the extent that they're influenced by them, the reflective scientist will regard that as illegitimate and will try to correct for it. If your symbolic system happens to be influenced by the language you speak (or common sense), and if you happen to notice that, then you'll try to overcome it, because it shouldn't be. The goal of the scientific language is something else – it's a constructed system aiming at a certain purpose, and the purpose is to capture reality by having symbols that designate, and predicates that have predicational properties, and so on. Natural language has nothing like that, and we can see it very clearly when we take the parts of natural language in which we talk about technical

operations and notice at once that we're not using the words in the scientific sense, even if we happen to be scientists.

I went too fast when I said that Quine had no argument for distinguishing the willingness to accept the form of 'gavagai' from the unwillingness to accept its meaning. He had an argument – meaning holism – which can take various forms. The weaker forms – the relativity of meaning to sentence and to language – are insignificant. Those are just misleading tricks, because the same arguments would give you sound holism, which doesn't confuse anyone. It would just tell us that a particular I-language has the property that sounds and meanings are correlated in some specific way and not some other way, and if they're correlated in a different way, it would be a different I-language. That's a triviality. So, there's nothing to meaning holism at the lower levels.

With regard to theory holism (the strong form), the one that Quine is really interested in, it's very hard to even state the concept. It's just kind of stated in a hand-waving fashion, and when you try to state it, it becomes quite hard, and you get into these questions like 'What does it mean to have the same word in different languages?', etc. Furthermore, it's based on an entirely illegitimate assumption that scientific language is part of natural language, and that's based on a misinterpretation of scientific language which takes things like "giraffes evolved long necks" literally instead of recognizing that they're shorthand for something else.

As to the problem about Bohr and intelligibility, Putnam surely has a point. We don't want a theory of science which leads us to the conclusion that Bohr was doing astrology. Nevertheless, however interesting a theory of intelligibility in science may be, it's not a basis for a theory of language, it's some other question.

Furthermore, as far as I can see, there's a perfectly good internalist interpretation of the Bohr problem. It makes perfect sense to say that all of Bohr's statements were false, literally, but he wasn't doing astrology. Although the statements were literally false, because there were no things such as atoms in his system, nevertheless, the structure of his theoretical discourse was similar enough to the structure of our theoretical discourse, so we can construct rather natural matches between them, and having done that, we can even take over a lot of his discoveries and make sense of an awful lot of what he was saying, even though it was all literally false. In the case of astrology – it's literally false and there are no matches that enable us to make it sensible. Therefore, that accounts, on completely internalist grounds, for the intelligibility of Niels Bohr without any strange assumptions about words

retaining meaning over time, the ancients having meant H2O by ‘water’, and so on and so forth.

Let me give a linguistic analogy (which is much simpler because linguistics isn’t as complicated as physics) that illustrates the point. If you go back to around 40 years ago, there was a debate in phonology between structuralist phonologists and what are called generative phonologists. They had a differing conception of what the basic unit was – the basic unit out of which phonological representations are constructed. According to (say) Jakobsonian structuralism, a phonological unit was a collection of distinctive features (where features are things that you can perceive, that you can get phenomenal/acoustic correlates for). The generative phonologists were saying, no, a phonological unit is an element in an abstract generative system which appears at a certain formal level. It doesn’t have features that you can hear. You may end up with things that have features you can hear, but that’s not what the unit is. They were having a debate rather like the debate between different theories of the atom.

Now, let’s assume for concreteness that the debate has been resolved in favour of generative phonologists. Does that mean that the structuralist phonologists were talking about units the generative phonologists were talking about all along? Does it mean that Jakobson was really talking about elements at an abstract level of some phonological representation? If you had asked him, he would’ve said, ‘Of course, not’, and he would be right because he wasn’t. He wasn’t talking about what nowadays people call phonological units. Was he talking gibberish? Of course, he wasn’t talking gibberish. In fact, we can make a lot of sense out of what he was saying, because there’s a close enough match between generative phonology discourse and structuralist-theoretical discourse to carry a lot of it over and reinterpret it and do it in unique and illuminating ways. That doesn’t mean that the discovery is lost. Maybe everything that was said was literally false, but the discovery is by no means lost – they just undergo a certain translation into a somewhat different conceptual framework.

Now, any linguist looking at this would surely say that this was what was going on. They would not say that Jakobson was really talking about contemporary phonological units.

When I was doing phonology 25 years ago and when I referred to something as a ‘phonological unit’, it was different from what people mean now. But I wasn’t referring to what they mean now, I was referring to what I thought then. That doesn’t mean I was doing astrology. Maybe some of it gets translated into modern terms,

maybe it doesn't. And there's no special reason why it should be any different in the case of the atom.

Before Avogadro, people used the word 'atom' and 'molecule' interchangeably (I think), so, are we really going to have to say that they meant by 'atom' what we mean by 'atom', and they meant by 'molecule' what we mean by 'molecule', even though for them these words were interchangeable? That's what you'd have to say in these theory-holist externalist theories. Obviously, it doesn't make any sense. They thought there were small things around out of which the world was constituted, indeterminately called 'atoms' or 'molecules', and we don't think that now. Now, we think there are two different kinds of things, one of which we call 'atoms' and one we call 'molecules', and they are kind of descendants of their earlier concepts, but descendants only in the way in which a contemporary giraffe is a descendant of an earlier giraffe – changes have taken place along the way which allow for certain matches but not others.

I'll come back to this in connection with theories of reference. I want to get back to the conclusion that contemporary theories of reference completely misconstrue language, because they're based on a notion that doesn't exist, namely, the notion of reference that holds between words and things. All theories of reference since Frege just don't seem to be true, although they may be true as the goal for scientific languages, which is another question.

The context in which these issues usually arise is in the context of 'analyticity', and that's just for historical reasons – because it's an interesting old philosophical notion, and because it's the one that Quine focused on when he was developing the theory. So, if some sentence is true whatever the facts may be, it's 'analytic', if it's true just because of the facts, then it's 'contingent', 'synthetic', or something like that, and then there's a big discussion about what the meaning of all this is.

Now, it follows from the theories of meaning indeterminacy, meaning holism, etc., that there are no analytic statements, because a statement that's analytic is true by virtue of the meaning. The statement "If John is dead then John is not alive" would be held to be analytic by virtue of the meaning of the terms, meaning you don't have to carry out an empirical inquiry into the world to find out that it's true. You know it's true because you know what the words mean, that's the idea. But, of course, if words don't have meaning, it can't be that the sentence gets its truth value by virtue of its meaning, obviously.

One of Quine's startling conclusions back in his early articles in the late 1940's was that there's no difference between analytic and synthetic sentences – they're all epistemologically the same. Some of them may be more deeply rooted in our belief systems or something, but nothing else. Since that's a question of fact, we could, presumably, carry out an empirical inquiry concerning the topic – we could try to find out whether, in fact, that's the way people understand sentences.

So, let's take a slightly more complex example. Suppose we give people the sentence "John persuaded Bill to go to college" and tell them that this sentence is true, and now ask for a judgment about whether the sentence "Bill intended to go to college" or the sentence "John intended to go to college" is true. Well, everybody knows that unless Bill in fact intended to go to college, John did not succeed in persuading him to go to college. But as to whether John intended to go to college or not, we just don't know, because the sentence doesn't tell us.

Those are facts and they have to be explained, so how do we proceed to explain them? There is a debate about this in the literature. Somebody says that there's a meaning connection in one case and just a factual issue in the other. Somebody else comes back and says, that's not my intuition, my intuition is that they're both factual. And then the debate stops. It's kind of like an impasse.

But it's not an impasse. Everyone agrees on the facts, and if you agree on the facts, then you can compare alternative theories (and there are alternative theories) on their merits.

The theory that gives you the obvious result – that says that "John persuaded Bill to go to college" entails "Bill intended to go to college" – is a theory that will involve notions like causation. The theory will, in fact, say that 'persuade' is a causative verb which has a lexical decomposition into a notion of causativeness and a notion of intention, so 'persuade' in effect means 'cause to intend'. And then people will look at some other language and say that 'persuade' is spelled out like a causative verb and has the properties of causativeness and intention. And then notions like cause, intend, agency, etc., will appear within a semantic theory about how concepts are organized and what their connections are and so on. And then there will be syntactic properties about causatives that we'll have to carry over to this explanation, and so on. That's one theory that gives you the conclusion that if John persuaded Bill to go to college, then Bill intended to go to college, but nothing about whether John intended to go to college.

Now, what's the alternative theory that we're supposed to compare with this? It's interesting that the theory doesn't exist. What we're told is that some sentences are more deeply embedded in our belief system than others. But obviously, that's not going to work here. The sentence "If John persuaded Bill to go to college, then Bill intended to go to college" is not deeply embedded in my belief system. In fact, I don't have any belief system involving that thing at all, or if you think that it is deeply embedded in my belief system, then show me your theory of belief fixation which leads to that conclusion. What theory of belief fixation shows me that one of those two conditionals is more deeply embedded than the other? Well, there is no such theory of belief fixation. Or some other people, for e.g., Paul Churchland, say that the difference is that what we call analytic sentences are deeply involved in normal inference and logic than others. Well, neither of these sentences are involved in normal inference or logic, so that can't be right. There is no alternative theory to the semantic theory which explains the difference as an analytic-synthetic difference in terms of notions like cause, intend, agency, etc.

We now have a funny situation. We have a class of facts that everyone agrees on. We have one theory to explain them, a semantic theory with semantic features with fixed meanings and so on. We have no other theory, just hand-waving. And, remarkably, everybody accepts this non-theory. What has swept the field is the non-theory. The standard conclusion that's drawn from these disagreements is that it has been shown that there aren't any analytic sentences. That's an extraordinary case of irrationality. However good you think this theory is, at least it's a theory – you can extend it, you can find evidence for it, and so on – but it's universally assumed that this theory has been disproven by the statement that another theory (which I can't construct) is right. That's, in fact, what has happened.

Therefore, it is now regarded as one of the best-established conclusions of modern analytic philosophy that there are no analytic sentences. The analytic-synthetic divide is artificial. Take a look at any textbook you like and that will be presented as one of the real discoveries of modern philosophy, as sort of like the bedrock for all further work. What has been discovered is that if you decide to abandon the one theory that works and claim that something else works, which you can't formulate, then you get this conclusion. Again, a rational naturalistic approach would draw exactly the opposite conclusion.

Now, of course, we haven't proven that there are analytic sentences. In the sciences you never prove anything (you haven't proven spacetime to be curved either, you just have evidence for it). What we've shown, however, is that there are pretty good semantic theories (just as there are pretty good phonological theories)

that explain a lot of the things that everyone agrees are facts, like the radical difference between these sentences, and it assigns them the different status ‘analytic’ and ‘synthetic’. Therefore, on naturalistic and rational grounds, that’s what we conclude.

To conclude on the basis of this that the analytic-synthetic distinction has been eliminated, or even questioned, is totally irrational. The distinction hasn’t been questioned at all. There’s a perfectly good theory that in fact predicts the distinction, there’s nothing that questions it, and therefore it hasn’t been questioned. Until somebody comes along with an alternative explanation of these facts, it hasn’t been questioned at all.

Again, that’s a case of extreme irrationality to insist that questions about language and mind simply can’t be approached by the rational requirements of the natural sciences, wherein if you have a fairly good theory to explain certain facts and a non-theory as a competitor, then the non-theory is thrown out and the fairly good theory is tentatively accepted, which in this case means accepting the analytic-synthetic distinction as pretty well-established.

Lecture #6

DURING THE DISCUSSION section this morning, some questions and comments arose about a quite interesting topic – about the nature of the triggering and the stimulus requirement (the input requirement) for acquisition of language, which bears directly on this so-called innateness hypothesis, which, as I tried to make clear, is not a hypothesis but just a truism, the question is just what form it takes. I mentioned one case, which is the extreme case of language acquisition in the case of zero input. However, what was required was social interaction – that was necessary. That's apparently true for other kinds of development as well – institutionalized children may be fed perfectly well, but they may not grow properly because of the lack of "proper" social interaction. This has even been shown experimentally with animals. So, if lambs are taken away from their mothers, they can get the same regular nutrition but they don't develop properly. Somehow, the effect of social interaction seems to be a trigger for mammals for getting things going, and the interaction among these three children, in the case I mentioned this morning, was, apparently, sufficient to get the language faculty developed on the basis of no input, but with full perceptual capacities.

There's another case where the perceptual capacities as well are radically impaired. This is an interesting case, and it was the one that was studied by my wife. In the era where people used to get spinal meningitis – this disease in early childhood which is completely debilitating – there were a fair number of children who were deaf and blind from very early ages, and it turned out that a number of them were able to learn language fully by a technique called 'Tadoma', which involves placing the person's hand on the face of the speaker, which is just a ridiculously limited amount of information. On the basis of that input and with training, people manage to attain what is so close to full linguistic capacity that you have to investigate exotic constructions to begin to find places where their knowledge breaks down. In normal interchanges, you wouldn't find any cases where their knowledge breaks down.

There was a project on this at the MIT electronics lab where they were working on sensory aids for people with sensory deficiencies, and one of the subjects in this investigation was a man who lived (I think) in Iowa with his wife, who was also deaf-blind and had learnt the same technique. The problem that they had around the house was to how locate one another, so there were various devices around that indicated where the other one was. But they could speak to each other just by keeping

hands on each other's faces. This guy had a job – he was a tool-and-die maker – which he held down. He flew to Boston alone. He carried a card which said "I am deaf-blind. May I put my hand on your face?" if he was lost or something. He could talk, it wasn't perfect, because he couldn't monitor his own speech production, so it tended to degrade and he had to get continued up-training to keep it up, but it was intelligible, and he could understand what anybody else was saying, and he lived an almost normal life.

There are a fair number of cases like these, and they show, quite remarkably, how little information is required to get the whole language system operated. Normal language-use is a dramatic enough example of the absurdity of believing that the language is taught, but these cases show with extreme strength that it's just nonsense to think that instruction from the outside determines the major course of language growth. At the very most, it can provide very peripheral information.

In the matter of the critical period, it turned out (for the population of people they were dealing with) that there was success in teaching language by this Tadoma method only if the deaf-blindness had occurred after about 16 months, meaning that the children weren't speaking yet but had exposure to language, and somehow that exposure must have set something in motion, which could then be built upon by the course of this training. The same was true of Helen Keller – the famous case of deaf-blindness – who became deaf-blind after an age roughly like that. It may be that being deaf-blind from birth would kill off those mechanisms, very much the way that (say) blindness in a kitten kills off its visual mechanisms.

A still taken from the 1953 movie 'Helen Keller in Her Story' illustrating the Tadoma method. Credit: afb.org

In the face of the general evidence about language growth (of which there's plenty), it takes considerable dogmatism to believe that conditioning is a factor in language growth. As more cases of this kind come along, it may be possible to sharpen and narrow down quite precisely the kinds of stimulation that get the system to operate and the kinds of sensory inputs that are required to get it moving.

There's now work by people who are involved with sensory aids that may shed some further light on this. There's a thing in the ear called the cochlear membrane which receives signals and is stimulated depending on the frequency of the incoming sound. There is an obvious idea that people are trying, which is to make an analogue of the cochlear membrane and simply put it on some other part of the body, and construct the analogue such that it would provide stimulation exactly in the manner the cochlear membrane would to the auditory nerves, and see if the human nervous system is flexible enough to adapt to signals coming from some other part of the body that have the same structure as the signals that tell the auditory nerve that it's a sound with a certain structure.

If this works, which nobody knows, it would be possible for a deaf person to hear with a device implanted on some other part of his body. There's a little evidence that something is working out, but nobody knows much about it, and, doubtless, there will be more inquiries of that sort.

Anyway, let's continue on the course that we were following. We were talking about the question of unification, which is an obvious question for a naturalistic inquiry. We have these studies of the mental systems (for e.g., computational-representational theories of language), there are cells in the brain, and, presumably, there's some relationship between what the cells are doing and what these systems are doing, and if we accept the plausible view of the post-Newtonians (people like La Mettrie and Joseph Priestley) that thought (language) is just a property of an organized system of matter (so, in this case, the generative procedure, the performance systems, etc., are some complex property of some arrangement of cells), then the question is, how do you fill in the blanks in that description? That's the unification problem.

I mentioned before that there are some connections here and there. The most surprising ones are the correlations of categories of linguistic structures with electrical activity of the brain (ERPs). But there are huge gaps. I mentioned a few ways in which people try to overcome their feeling of discomfort with those gaps. Notice that you don't have to share those feelings of discomfort. There's no particular reason for discomfort. There are lots of things we don't know, and one of them is how to fill in this gap – not more shocking than other things. But a lot of people feel there's a special problem here – something involving the mental.

Among the ways of trying to overcome that feeling is the slogan 'the mental is the neurophysiological at a higher level', which is just dogmatism, because we don't know that it is the neurophysiological that's relevant, and there's no evidence about it. I suggested that the slogan ought to be rephrased as 'we might speculate that the neurophysiological is the mental at some other level'. And maybe it is, maybe it isn't.

Another approach to overcome feelings of discomfort is what's called eliminative materialism, that we should stop studying the mental altogether and study the material world – that will become a substantive proposal when somebody tells us what the material world is, and short of that it doesn't mean anything. The mental is part of the material world since there's no other world. An associated proposal is that psychologists and linguists should stop studying cognitive processes, and language, and so on and start studying neurophysiology, which is just craziness.

It's like saying that embryologists should start studying quantum physics, which makes no sense, of course.

Another view is that we should try to think in terms of connectionist models (which I loosely described, but their important point is that they don't have any explicit rule-following in the way that an ordinary computer program or a computational system does) and maybe somehow they'll do something. Again, that's pretty irrational. There's no reason to believe that these abstractions from the nervous system will select the right properties, or whether they have the right kind of structure, and there's no explanatory success in using them.

The major approach, the one I want to turn to now, is the idea that we don't really have to feel uneasy about the mental because we have good, hard-headed, solid, substantive examples of robust objects that have mental properties, and nobody worries about those, therefore we don't have to worry about humans, and these robust objects are called computers. The idea is that modern computers are physical objects – everybody agrees that your PC is a nice robust solid physical object – and it does execute software, and if all that mental processes are is software, then a physical device like the computer is doing it, so there can't be anything mysterious about the mental, and now we've overcome the problem.

The view that psychology is the study of software problems was coming forth in about the mid-1950's, and it was certainly part of the so-called cognitive revolution, and in some respects, it was kind of a liberating idea. You have to recall that this was a period when psychology meant the study of things like conditioning (which, probably, doesn't even exist) or learning theory (which was a total fiasco, in that it achieved almost nothing in a century), and the various branches of psychology were more or less designed in terms of one or another dogmatic assertion about what kind of a theoretical posit you're allowed to make, which is total nonsense. So, the field itself was totally stultifying, and to come along with an idea that allowed some work to be done was exciting. Hence, the idea that the study of cognition is really the study of software problems made it possible to open up a lot of topics for inquiry, and that was what led to the cognitive revolution. In retrospect, it's less obvious that this was such a great move, because it brings along with it a number of serious conceptual problems that don't arise if we just approach the subject along a straight naturalistic path.

The first problem that arises has to do with the nature of artifacts – things that are constructed. Take, say, this thing over here. If somebody asked me what it is, I'd say a 'table' or a 'desk'. Am I right in saying that? Well, in order to answer that

question, it's not enough to do a full investigation of this thing. You can understand every physical property of it and not know whether it's a table. Maybe I think it is a table, but it's really a hard bed for a dwarf who comes in here every night to sleep on it, and that was the designer's purpose, and that's the way it's normally used, and it just happens that I walked into this room by accident and mistook it for a table and so I'm sitting here. That's perfectly possible. If that was the designer's intention, then that's the standard use – it is a hard bed and not a table. So, whether it's a table or a hard bed or some other thing depends, at least, on the intention of the designer, on the customer-use, on the functions that it serves, and so on and so forth, that is, what it is depends on how it's placed in a very complex framework of human concerns.

Now, no physical object can have those properties. It's not that this table isn't a perfectly fine physical object, it's just that what individuates it as a table or a bed or something else is relative to specific properties of human cognition and has nothing to do with its nature. A Martian lacking our concerns and interests and knowing all the physics we knew, would know everything about this that we do, but wouldn't know that it's a table, and, in fact, couldn't even ask the question, because he wouldn't know about things like intent and use. And that's true of all artifacts, but it's not true of natural objects.

So, whether something is a nematode or not is independent of the designer's intent – because there's no designer. It's independent of its function – because it doesn't have a function (maybe you use it to poison people, which doesn't affect whether it's a nematode or not).

So, there is a basic difference between objects of the world and artifacts. This was well understood as far back as Aristotle. If you look at Aristotle's classification of objects into *aitia* (it's usually translated into English as 'causes'), meaning the factors that enter into making something what it is, they include intent, function, structure, constitution, etc., and what an entity is depends on how it is placed in this framework. Now, Aristotle thought about this as a metaphysical analysis, but we might prefer to think of it today as an epistemological or a cognitive analysis, that is, an analysis of how human intelligence organizes the world for its own purposes, but the story is about the same. In fact, one of the (in my opinion) best theories of semantics that's around was suggested by a philosopher named Julius Moravcsik, who described it as the 'aitiational theory of semantics', meaning a theory of semantics that takes the Aristotelian factors and uses them as semantic categories, suggesting that Aristotle was just intuitively picking out the properties that humans used (like function, purpose, intent, etc) and categorizing objects that way, and that's

the core of human common-sense natural language semantics, and he gave a series of examples that illustrate that that's plausible, and I think it is plausible.

Actually, it's kind of striking that if you look at contemporary work in philosophical essentialism that grew out of Saul Kripke's *Naming and Necessity* (in which he gives arguments (which I don't think are convincing) about the essential nature of entities), it turns out that his essences are virtually identical with the Aristotelian aitia. So, *being* a table is the essence of this thing, that it's an essential property of this table, in Kripke's essentialism. That just seems to me a very misleading way of saying that, as human beings, we categorize things in terms of a space of human interests, which includes things like the designer's intent. It doesn't tell you about the essence of this object – its essence is just whatever it's made up of. It's being a table is something about our interpretation of the role it plays in our lives, nothing more than that.

This, incidentally, is why the natural sciences are obviously going to depart very far from human common sense. The natural sciences are not going to have in them any notion like 'being a table', which will hold of it or not hold of it depending on its designer's intent – that's not going to be a concept of physics, for example.

Now, this is true of every artifact. It's also true of artifacts that are designed to do something. Take, say, a key that opens a door. If I ask for a key to the door of the room of the hotel I'm staying in, and if they give me a thing, and I try it and it doesn't work. Did they give me the wrong key? Did they give me a key that malfunctions? Did they give me something that wasn't a key at all? There are no general answers to that – it depends on the purpose for which it was designed and for which it was given to me. If the purpose of the guy behind the desk was to cause me problems, then the key is functioning perfectly when it doesn't open the door, it's not malfunctioning. You can't tell whether an artifact is malfunctioning just by looking at what happens when you use it, you have to place it in a much richer framework of human concerns and intentions, including things like designer's intent (or, in this case, intent of people who are involved in providing it), and that's true for all artifacts, including, of course, computers.

Coming to computers, they are artifacts. They are designed to do something. Suppose somebody gives you a software program that adds numbers, and suppose what it adds is what Kripke calls the 'quus' function, which is different from plus – it does what plus does for everything except for some two numbers, for which it gives the wrong number, maybe 5 x 25 = 139 or something, and the rest of the time it gives the right numbers.

Well, is the program malfunctioning? That depends on what the intent of the software designer was – it's functioning quite perfectly if he wanted you to get the wrong answer when you multiplied 5 x 25, or it's functioning perfectly if he had another function in mind that didn't happen to be plus but just happened to be identical to plus in all cases except this one. There's no answer to the question whether it's functioning properly or following a rule or whatever. It's just doing what it does, and we interpret it as following a rule, not following a rule, functioning properly, malfunctioning, etc., depending on the space of intentions, goals, purposes, etc., in which we place the object. But the question of whether that software program is failing to follow the rule of multiplication is a meaningless question – it's failing to follow the rule only if it was intended to follow the rule, but there's nothing in its nature that tells you whether it's failing to follow a rule.

That's going to be true of computers quite generally. Furthermore, in the case of artifacts like computers (or keys or desks), there is no natural kind. There's no category of nature into which they fall. Just about anything could be a computer. A pile of sand on a beach could be a computer if you give it the appropriate interpretation, like, if you interpret the arrangement of grains in such a way that it's calculating numbers. Anything could be a key for a properly shaped door (this table could be a key for some door that was made in a crazy fashion). There's no normal case for artifacts from which other things are deviations, which is quite different from a natural object.

In the case of a nematode, a cell, a giraffe, etc., there's a natural kind, and there could be a word at least in science (but not in human language, in my opinion). In some created scientific language, there could be a term that picks out that natural kind, like 'nematode', and there could be normal instances of it, and it would make sense to talk about deviant instances – for e.g., it would make sense to talk about a nematode with a genetic defect (maybe it was hit by a cosmic ray or something), because there is a normal nematode, and we can therefore make comments about the category of nematodes, but we can't make comments about the categories of computers, since anything can be a computer (if you have enough imagination).

Similarly, if you look at a computer, you can't really tell what's the hardware and what's the software. It's not like some of the chips are labelled 'software'. What's called software could be distributed around the computer chips in any crazy fashion. If you consider only certain kinds of modifications to the computer, like not hitting it with a wrench but sticking a disc in a disc drive, and if you restrict yourself to one category of them (sticking discs in disc drive), then relative to those quite arbitrary decisions (about what kind of interferences with this object you're allowed

to make), you could distinguish software from hardware. But that's again relative to human concerns and interests, the computer itself doesn't know anything about hardware and software.

So, when we move from the study of the mind (which is some property of organized matter) to computer models, we've moved out of the frying pan and into the fire, to use the standard English idiom. You've moved from a problem that's manageable to something that's completely unmanageable, which introduces inherent complications that don't arise in the case of the study of thought in the first place.

Now, the belief that there was a problem about the mental was a mistake – there was no inherent problem about the mental, other than our lack of knowledge about what it is, but that's a problem that holds all throughout the sciences, so there was no special problem. So, that was a mistake in the first place. But this effort to resolve the problem just adds new difficulties that didn't arise in the first place.

In the case of language, it is a natural kind object (just as nematodes are), and it has a normal case. It's not a matter of our decisions whether rules are being followed or not. It's just a fact about the system (like it's a fact about our visual system that it's processing a certain way). There is no designer's intent because it's an object and not an artifact. There's no conventional use because there's no use at all. It has no primary function because it has no function – things in the world have no function, we assign them functions because of the way we choose to look at them. Hence, it raises none of the problems computers raise. Therefore, using computers as a robust model to make you feel happy about the mental was, from this point of view, a mistake to start with. It's a mistake that shouldn't have been made, and I think it's a mistake that arises only if you fall prey to dualistic temptations – temptations that lead you to feel that there's something special about the properties of organized matter that we call thought – and that does not arise in the case of the properties of organized matter that we call attraction and repulsion.

Moving on to the cognitive sciences, let me make the distinction between the empirical studies (like the study of visual processing) and the theoretical inquiries into the nature of the subject (the more philosophical side). In the latter kind, some rather curious moves had been made which follow from picking computers as the model and thinking of the mental as software problems.

Notice that you're led very quickly to a conception of the mental as being whatever that can be computed. As it is put in the AI literature, mental activities are

just the carrying out of an algorithm. And an algorithm is just a computer program, so mental activities are the carrying out of a computer program (in fact, many advocates of AI argue that thermostats have mental activities because they're carrying out a certain algorithm). And computers have mental activities, or they would if they have enough achievement – it's often suggested that they have to achieve a certain amount before we're willing to attribute to them mental activities.

This gives you a different approach to the mental. In fact, this is another form of dualism – it's treating the mental in some fashion quite different from the physical world – and it leads in strange directions. Instead of thinking of I-language and properties of I-language as states of the brain (as mental organs, if you like, which are just like any other organ except that they happen to be aspects of the brain), what you do is totally move over to the computational level and consider what can be done by a computer program.

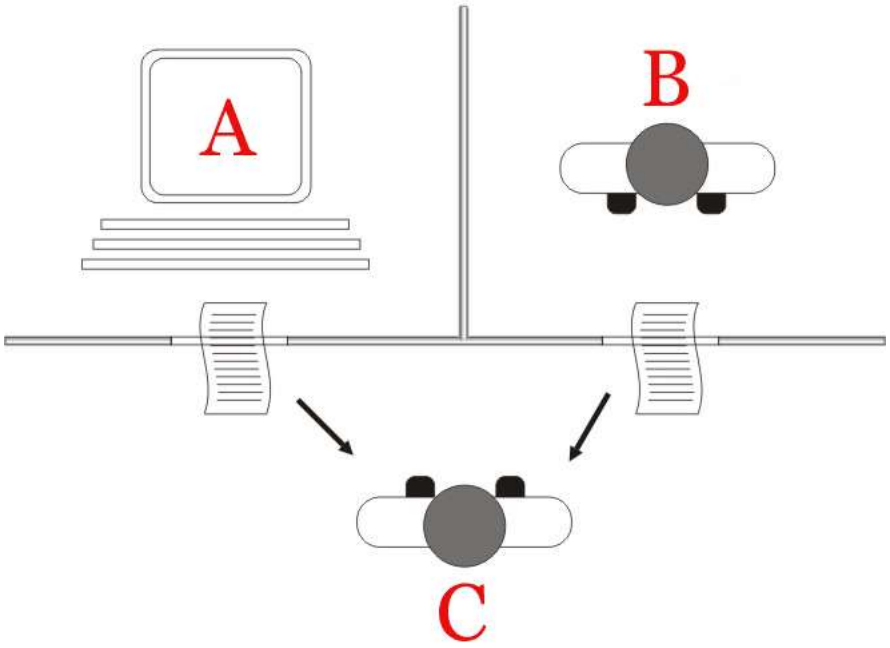
Now, if you're not inclined to say that thermostats have mental properties (and many people are not), you'll add certain performance criteria and you'll say that an algorithm has mental properties only if it satisfies such criteria, and the standard performance criteria, the one that sort of dominates the literature, is what is called the 'Turing test'.

Alan Turing was an important mathematician of the 20th century who, among other things, did some of the mathematical work that led to the design of modern digital computers, and back in 1950 he wrote a classic article in which he proposed what has since been called the 'Turing test'. In my opinion, his article has always been misunderstood. I don't think people have read it correctly or have not read it at all. I'll come back to what he actually said, but let's begin by what he's thought to have said, which is the only thing that influences the field (and not what he said).

Firstly, the words 'machine' and 'program' are interchangeable. Any computational type of a machine that somebody builds is a program for what's called a Universal Machine (UM), which is not very exotic – if you have a PC then you have a UM, with one proviso – you have to be able to stick an infinite number of possible disks in it to keep building up the memory. But if you've got an infinite source of additional memory instead of just a finite source, it becomes a UM (in fact, it becomes what's called a Turing machine), and any program for such a machine is a possible machine on its own, and any possible machine on its own is a program for UM (that's what it means for the thing to be universal). So, any possible algorithm that exists is a program for your computer if you idealize it to infinite memory –

that's basically what it comes down to (it's not exactly accurate but good enough for our purposes).

Turing designed the following test in his paper:



Credit: commons.wikimedia.org

You have a person who is a judge (C) in one room, and you have another person (B) and a machine (A) (which just means a piece of software with a computer that runs it) in two other separate rooms. The judge is allowed to communicate both to the person and to the machine in the same way (say, by typing something into the keyboard that the other two pick up at the other ends). The judge is allowed to ask questions both to the person and the computer, and then the person types in an answer and it comes back to the judge, and the computer program does whatever it's designed to do and types out a response. And if the judge can't tell which is the person and which is the machine (if the judge is confused, in other words), in that case the machine has passed the Turing test. And what Turing is alleged to have said is that if a machine passes the Turing test, then it thinks – that all human intelligence is is the capacity to pass the Turing test, that is, to fool somebody into thinking you're a person. Like, if a Martian comes down and fools us into thinking they're a person, then it thinks.

Notice that this is a very non-naturalistic idea. This is a radical shift from (say) La Mettrie and Priestley, for whom thought is a property of matter on this earth (it's

a property of particular forms of organized matter just as electricity is a particular property of certain forms of matter). The fact that you might be able to create something out of non-earthly matter that has properties like electricity has no bearing on what electricity is. The idea that you might be able to fool somebody into thinking that something is a duck doesn't tell you it's a duck, because to be a duck is to be a property of a certain kind of matter. But this is totally different.

Now, the mental is regarded first as computational, and then if you're unhappy about thermostats being mental, the computation will be sort of sophisticated enough – where 'sophisticated enough' is spelled out in terms of passing the Turing test. That's a completely different conception of the mental, and the mental now becomes a property unlike any property known in the natural world – the mental is defined by a certain performance criterion that's stipulated, and any class of objects built out of certain kinds of components that passes that performance criterion has this property. There are no such properties in the physical sciences. If you think through physics, chemistry, or something like that, there's no category of entities identified in terms of them satisfying certain performance criteria. You may use performance criteria as tests for its own property, like, you may use a litmus test for determining whether something is an acid or a base, and you may rely on that test enough so that you may start calling it a definition, but the fact of the matter is that being an acid is not defined in physical theory as passing a litmus test – it is whatever it is, and it happens to pass the litmus test. Performance tests don't exist in the sciences, they are just an operational criterion. So, this is something entirely new and entirely non-naturalistic. Again, I'll qualify by saying that it's not clear that this is what Turing was saying, but this is what he is interpreted as saying.

Notice that when you talk about the Turing test, that's misleading, because what I just described is not a test but an indefinite range of possible tests depending on exactly how you set it up, and there are all kinds of different ways of setting this up. So, the so-called Turing test is a class of possible performance criteria. And remember the course we followed. We start from feeling that there's some problem about the mental because we can't explain it in terms of cells, and that this is different from the one about chemical elements in the 19th century when you couldn't explain it in terms of particles and fields. We try to resolve the problem by saying we have robust objects around that have mental properties, viz., computers, and so it is nothing mystical or ghostly. We then identify the mental with the software aspects of the computers, overlooking the fact that that's not a physical distinction – it's a distinction we make in terms of a class of interests and concerns. We then identify the mental with the computational, or with a sufficiently sophisticated computation.

It now turns out that when we're studying the mental, we don't really care what the physical realization is at all. Now we have a distinction between the mental and the physical – the mental is just abstract computations, and the physical form in which the computation is carried out is irrelevant to the study of the mental, because we're just studying certain algorithms that can be carried out any way we like, and we're assuming that there really is an identifiable category of software vs. hardware in computers independent of our interests and concerns, and we're assuming that it is really possible to pick out in the world what is a computer. I don't agree that any of this is possible, but to continue with the discussion, let's grant it.

Well, we now study algorithms, viz., the mental, in abstraction from any form of material realization, and that becomes psychology. Cognitive psychology is now the study of algorithms, in fact, that's the way it's usually considered. I'll give some quotes from a particularly lucid explanation of it from Ned Block (a colleague of mine), from his chapter on psychology taken as the computational theory of the mind in the three-volume encyclopedia of the cognitive sciences published by MIT a year ago. He says, "the computer model of the mind is a level of description of the mind that abstracts away from biological realizations of cognitive structures." Here we have to be a little cautious, because there are a lot of different ways of abstracting away from something, and we have to make sure which one we're talking about.

For example, abstraction away from something may be what we may call 'contingent' abstractions, that is, we may study something abstracting away from some of its properties because it's useful to do so or because we don't know enough about the other properties to look at them. We do contingent abstraction for a particular purpose or because of a particular lack of knowledge. For example, you might decide to study the planets as mass points following certain laws, abstracting away from the questions of whether they're made out of atoms or whether they have mountains on them or whatever. That's perfectly intelligible, and in this case, it leads to a certain field of mathematics called rational mechanics that you can investigate because it teaches you something about the planets. So that's one kind of abstraction.

Another kind of abstraction that we might consider is 'exploratory' abstraction – we take some system too complicated to study, and we identify certain properties of it, and we just study those properties and say that maybe if we study just those properties then we'll learn something. So, you abstract away from a full complexity just to explore, and that's done all the time. In fact, every experiment in science does that – you have something messy, and you abstract away from most of the mess and look at particular things (in fact, you design experiments so as to yield

only those particular things), and you study them in the hope of learning something about the world. That's exploratory abstraction. Both these abstractions fall within normal naturalistic inquiry.

However, there is a third kind, and that's the kind that's proposed here – what we might call 'principled' abstraction, i.e., it just doesn't matter how the systems are realized in mechanisms. We're abstracting to something where it just makes no difference what the physical realization is. You might say that this is done in pure mathematics. If you look at Euclidean geometry in the classical period, was it a study of the physical world or was it mathematics? For a long period of time, that wasn't a clear question, that it was sort of both. By the time mathematics got sophisticated enough, it became a clearer question, and it even turned out in fact that Euclidean geometry is not a good theory of the world. But when you abstract Euclidean geometry away from the world entirely and go to Hilbert's axiom system and study its properties, then you're doing pure mathematics, and that's principled abstraction. But, of course, that's not science at all. It's pure mathematics, and it may be useful for science, like a lot of other things are, but it's not science and nobody thinks it is. So, there's nothing wrong with principled abstraction. In fact, one might think of large areas of mathematics as principled abstraction. But here we have something new – principled abstraction in an empirical discipline. And the reason for these side remarks is to make clear how new it is.

So, in the study of the mental (the computer model of the mind), we make a principled abstraction from biological realization. We proceed as follows. We assume a class of abstract machines and think of them as being made of certain elementary components called 'processors' (they're not piece of matter or anything, just abstract objects having formal properties). In practice, these processors are selected so they have what's called 'universal Turing machine capacity', i.e., they're rich enough for you to construct a universal Turing machine out of them. So, we pick these abstract processors and we then set up a performance criterion (like passing the Turing test), and we consider a class of objects – the set of all objects K where K is constructed out of P (processors) and satisfies C (criterion).

Let me put it simply. To identify the primitive processor, we just give its properties, and anything at all that has those properties is a primitive processor (whether it's made out of materials that exist in our universe, some other universe, or whatever, doesn't make a difference). We set up a performance criterion, which could be anything you like. Performance criteria are completely free for the asking. The ones that are usually looked at are some version of the Turing test. We then consider the class of machines (the class of abstract objects) that are constructed out

of these processors, that is, that can be put together out of elements of any kind whatsoever that have the properties of the processors and that pass the criterion. And that's psychology – the study of what I just identified.

Now, there's usually another thing added here, and that is that people have to be in K, and that's not so obvious, incidentally. It looks completely arbitrary, that you have to include people in K. It's not clear what they're doing in there at all. This is not mathematics (because there are no theorems or anything), but as an abstract inquiry, it doesn't seem to make a lot of sense to put people in K, but that's usually assumed tacitly – that the performance criterion is the one that *people* pass.

To continue with Block's description, for the cognitive sciences it does not matter how the primitive processors are realized. So, they can be realized in grey matter, in switches, in mice pulling strings that make something work, in other kinds of matter that don't exist but that we could imagine, etc. We don't care at all how they're realized, we only care about their properties, and we're interested in whether the software built out of those entities can pass the performance criterion, and if it does then it's part of K as long as humans are also a part of it.

Block then goes on to point out, correctly, that psychology is not a biological science. Furthermore, he says, given the anti-biological bias of this approach, if we can construct automata in our computational image (automata that do what we do, relative to a performance criterion), then we will naturally feel that the most compelling theory of the mind is the one that's general enough to apply both to them and to us, as distinct from a biological theory of the human mind which would apply just to us. That's the crucial difference. And he wants to insist on this because of the common though mistaken idea that psychology is a biological science, and according to this it isn't – it's not part of biology at all, therefore it's totally non-naturalistic. So, psychology or the computational model of the mind is now some new form of intellectual inquiry, divorced from the sciences, which is considered with certain categories of abstract systems that satisfy certain performance criterion, and for unexplained reasons humans come in at this point.

Note where this differs from the natural sciences. The natural sciences never abstract away in principle from the mechanisms. That's unheard of. Suppose the natural scientist had two theories of language, T1 and T2, and suppose he were to discover that T1 is realizable in human cells and T2 is not. For the naturalist, T1 would be selected and T2 rejected. But for the computational theory of mind, that would be irrelevant, because, to paraphrase Block again, it does not matter how the processors are realized – that is, for the theory of language, the potential discovery

that some property of human cells selects one theory over another would just not be relevant; language is abstracted away from that in principle. That's one striking difference between this and the natural sciences.

The second difference is that the natural sciences are not concerned about any kind of criteria – whether they're performance criteria as defines the whole field of AI or other kinds of criteria like structural resemblance and so on. Such things don't appear in the sciences. If something meets the criterion (whatever it may be), it's of interest for the sciences only insofar as it tells you something about the real objects under consideration.

Let's take the case of chess-playing computer programs, which, as Herb Simon points out, correctly, was the *Drosophila* of the cognitive sciences. Why was chess-playing the Drosophila of the cognitive sciences? The main concern around which the field evolved was chess-playing machines, because there's a particular version of the Turing test which says that a chess-playing machine passes it if the judge can't tell whether it's a grandmaster (another way you check that out is by seeing if the machine can beat the GM).

Notice how a natural scientist would look at this (assuming that they would want to study chess playing at all, as I mentioned the other day that it's probably not a reasonable topic to study). Suppose somebody comes along with a computer program that satisfies a chess-playing version of the Turing test – that it fools someone into thinking it does as well as (say) Kasparov. The first question the natural scientist would ask is – how does the program work? Second question is – how does Kasparov (or any person) work? Third question is – does the first question tell me anything about the second? – if it doesn't then I'm not interested in it. That's contingent/exploratory abstraction – you study some abstract object to see if it will teach you something about the real things, and if it doesn't, too bad. And, in fact, chess programs don't teach you anything about the world, so they're of no interest. The fact that they've been at the core of cognitive sciences around which all general research evolves just illustrates the strength of the Turing test in defining the field.

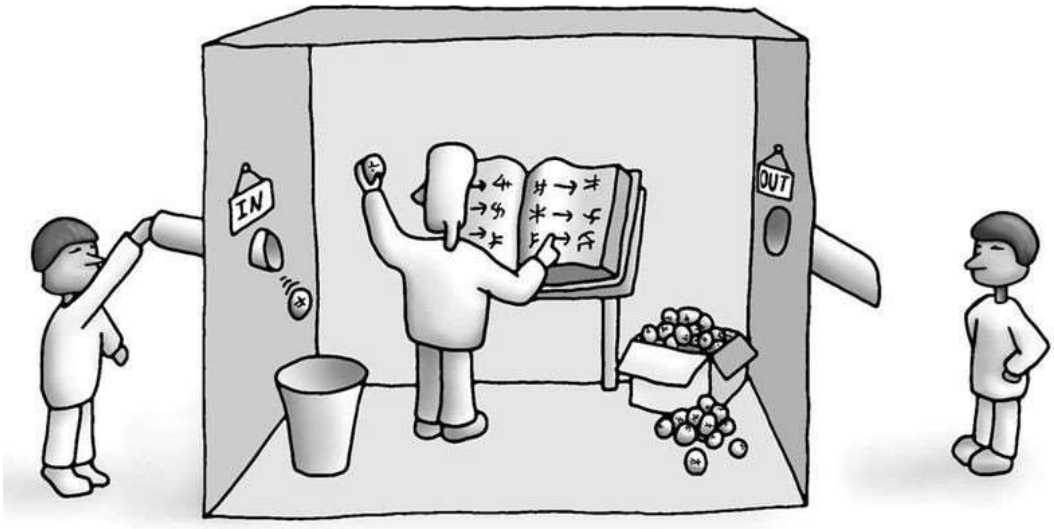
So, that's another sharp departure from the natural sciences. And, in fact, when you enter this non-naturalistic domain, the questions that arise are pretty odd. Here's a standard one that's been debated for years in philosophical literature.

I'll start by quoting from a classic paper by John Haugeland in *Journal of Philosophy* about ten years ago. He asks that how might one empirically defend the claim that a given strange object plays chess, and he gives the conventional answer.

Notice that it doesn't matter out of what it's constructed because we've abstracted away from biology in principle. He says that we would properly conclude that a given strange object plays chess if a skeptic is convinced that it meets a certain performance criterion (which he describes as some version of the Turing test). If the skeptic is convinced that the machine meets that criterion, then, Haugeland says, we would have empirically defended the claim that the object plays chess. The idea is that this is an empirical claim (which is a little odd because it is not a part of the natural sciences) that we're trying to defend, and we successfully defend it if the skeptic is convinced by the Turing test that the thing is passing the criterion.

The second part of his paper goes on to understanding language, which is what the paper is called – *Understanding Natural Language* – and he gives a bunch of arguments to say that it would never be possible to convince the skeptic that the device is indistinguishable from the person. So, the same framework is accepted, but we would've empirically defended the claim because the skeptic would never be convinced. Therefore, the machine is not understanding language. Notice the crucial point that this is an empirical claim, and the claim is correct if the performance criterion is satisfied – the same basic argument.

There has actually been a huge debate about this over the last ten years or so initiated by John Searle, who wanted to debunk strong AI, and he raised the following question – let's imagine that somebody constructs an algorithm that understands Chinese by whatever performance criterion you like. Suppose you make a room and you put John Searle (JS) in it, who by assumption doesn't understand Chinese. Suppose the test you're trying to provide is – you give a question in Chinese and you're supposed to give an appropriate answer in Chinese.



Credit: commons.wikimedia.org

Suppose you feed the question in, and suppose JS has lying in front of him a huge table where the actual algorithm is written out, and JS gets a series of lights flashing which are the representation of that question after it enters the system, and he executes the algorithm and types out the answer, and that fools the observer, by assumption. Searle then argues that though the answer is given, no understanding is going on – because JS doesn't understand Chinese.

Searle wants to destroy the whole computational theory of mind (like what Block is talking about, AI, and so on). Searle says that no algorithm can ever carry out a mental action, hence to identify the mental with the computational is just flat wrong, Turing test or no Turing test. And the proof of this is that if JS were able to carry out this algorithm, obviously no understanding went on, therefore that shows that no algorithm can be a representation of the mental process. Therefore, the whole computer model of the mind is wrong.

The answers to this are generally of the type that Searle's overlooked the question of speed – it would take JS forever to run through the algorithm but the machine will do it instantaneously. So, Searle comes back and says, suppose you add the whole population of India inside the machine, assuming none of them understand Chinese, and they're all racing around like maniacs doing pieces of the algorithms. And then somebody else comes back and says that the point is maybe that JS doesn't understand Chinese, but he plus the room understands Chinese, and then Searle comes back with an answer about that, and so it goes, up till today.

Those are the kinds of questions you very quickly get into when you make the non-naturalistic move that abstracts in principle away from biology. In fact, you get even stranger results than this. I'll bring them up tomorrow.

Lecture #7

I'VE BEEN COMPARING two different general approaches to the status of what is called the mental, which is just a descriptive category. To make it concrete, this would mean two different approaches to things like the creative aspect of language use, the problem of understanding Chinese, the formal relations of expressions such as rhyme and entailment, problems of things like anaphoric dependency (which is a core part of what might be called the logical structure of language), and comparable phenomenon in other domains of the mental generally. So, when we talk about the mental, we're talking about that category of phenomenon.

One approach (and the one that I'm advocating) is the naturalistic approach expressed by post-Newtonian scientists such as Priestley, who argued that these properties are just properties of organized matter. So, rhyme is a property of organized matter, just like the property of attraction and repulsion, electricity and magnetism, etc. are. And if you believe that, your next task is to spell out the properties and to develop a theoretical account of the properties, the states, and so on, and then to show how organized matter can have these properties – that's the unification problem – and if the history of science is any guide, that may very well require radical changes in the core parts of science, which has always been the case in the past. So, one naturally keeps an open mind, but in this case the gap between the properties and what is known about the brain is so enormous that the idea that there might well be radical changes in the core sciences necessary is not at all surprising. So, that's the naturalistic approach, and we know how to proceed with it – maybe it's hard, maybe it's easy, maybe we make a mistake or not, but there's an intelligible research project and people pursue it in various domains.

In contrast to this, there are various approaches that I've been calling 'non-naturalistic' and 'dualistic', meaning not metaphysical dualism but some kind of epistemological/methodological dualism. I discussed a number of such cases, and a lot of them converge around Quine's radical translation paradigm.

Yesterday, towards the end, we got into another one – the computer model of the mind – which has the basic conception that the mental is the computational, so the mental properties are just software properties. Well, a computer model of the mind could be completely naturalistic, that is, one could say that among the

properties of the brain are properties of executing certain algorithms. That's a perfectly intelligible statement. The concepts are all clear, and we could proceed to investigate it. But this computer model of the mind, crucially, does not take that step, and it insists on non-naturalism. It insists, crucially, on abstracting away from biological structures in principle – no temporary abstractions, because we don't know or because we're trying something exploratory. So, it insists that psychology viewed in terms of this version of the computer model of the mind is not a biological science. That's the crucial point.

The second crucial point in the way this conception is formulated is that we are interested in automata that are in our computational image – that are enough like us, more or less. So, our study is not a study of humans but a study of a class of automata that are in the computational image of humans, and, as far as I can see, we could just as well study other classes of automata that aren't like humans. That would be just as legitimate, since it's a non-physical, non-biological science anyway. I quoted this picture from Ned Block's article in the recently-published encyclopedia of the cognitive sciences. The same view, with various modifications, is developed all over the place. Another good source is the philosopher Justin Leiber's book *An Invitation to Cognitive Science*. He gives a similar picture. That book has the advantage of tracing things back to the first cognitive revolution, he knows the history, and he focuses much more clearly on Turing and the Turing test. So, that's the non-naturalist version of the computer model of the mind, to be distinguished from the naturalist version of the computer model of the mind.

Now, the non-naturalist version of the computer model of the mind – the one that prevails in the field almost entirely – does raise a certain number of questions. For example, if we have some (mathematical or non-mathematical) object X, we have to ask – when is X a computer? If this is just pure mathematics, we can stipulate it – it's a computer if I say it's a computer. But if this is supposed to be an empirical study (and apparently it is, because people who believe in it ask questions like 'How can we empirically decide whether a strange organism is playing chess?'), even though it is a non-biological subject (a curious idea to begin with), then the question 'when is X a computer?' is not easy to answer. That's a point that actually Searle has emphasized, correctly, for many years. Anything you like is a computer under some interpretation, and the things that are computers are not computers under other interpretations, and this is a general problem about artifacts that goes back to Aristotle. The question of when an artifact is a table, a desk, a key, a computer, etc., is a matter of intentions, its place in our space of human interests, and all sorts of other things, and it is not part of its nature. Notice that the question doesn't arise in

the naturalistic version of the computer model of the mind – the brain just is what it is. If one of its properties is being a computer then that's what it is.

Another question that has to be asked is that given that X is a computer, how do we decide that Y is its program? In other words, in a computer, we make a distinction between hardware and software, but the distinction is not a physical one, it's a distinction based on our interests and concerns. And, again, if this is pure mathematics, we can just stipulate. If it's a naturalistic version of the computer model of the mind, again, there's no problem – we just say it's true. Just as things have other properties, they have the property of executing certain algorithms, and that empirical claim has to be evaluated by the usual criteria of scientific inquiry – does it fit properly into an explanatory theory, and so on and so forth. But in the non-naturalist computer model of the mind, this too is a problem.

A third problem which arises whether you regard this as pure mathematics or not is – what counts as 'in our computational image'? What does it mean to be enough like us? Obviously, that has to be spelled out whether you regard this as pure mathematics or an empirical study or whatever. And the overwhelmingly dominant answer to that is – passing the Turing test, which, as I said, is not a specific criterion. It's just a range of possible criteria that you can set up one way or another, but it's an intelligible range of criteria. Basically, what this says is –something is in our computational image if it passes certain performance criteria, that is, something is in our computational image if it would be indistinguishable from human performance by some measure or framework. And that is Turing's view, with a modification I mentioned last time. The reason why Turing has commonly been alleged to have started off the field of cognitive science with his 1950 paper where this is proposed is that it is the standard answer to what it means to be in our computational image, and it's the answer that pretty much defines AI, most of cognitive science, and a wide range of philosophy of mind, and if you want a sense of how wide a range that is, just take a look at the huge literature on Searle's Chinese room argument and also look at the range of people who have tried to refute his argument (and they are all implicitly accepting the Turing test criterion, because that's the way his problem is set up). Incidentally, you might have some other criterion – to be in our computational image means to be enough like us *in some other way* – and still the same kinds of problems would arise. So, we can, for the purposes of our discussion, keep to this one, which is at least loosely characterized by Turing.

Several people pointed out after class that there are all kinds of ways in which a judge could trip up the computer, and there are. But these are all discussed in the literature. For example, if the answers come back too fast for a complex arithmetical

operation, you will know it's a machine, therefore to pass the test you'd have to slow it down. Or if it takes computations that require a hundred years, then it's a machine, therefore you set it up so that it collapses before the computations become too long. There's a ton of literature on this, and we can put it aside. We'll just assume that various precautions have been taken to overcome the obvious ways of deciding whether something passes this criterion. That's not where the real problems arise. The problem doesn't arise in sharpening up the tests. Where the problem arises is in – why do it altogether? That's a different question, and it doesn't matter how you fix up the tests.

Now, this third problem, which obviously has to be answered before the computational theory of mind has any content, doesn't arise in the naturalistic computer model of the mind, because the brain just is what it is. Period. We're not asking what's in the computational image of the brain, we're asking what the brain is – it's as if we're studying embryology and not asking about abstract systems which are similar enough to chicken embryos so that we study that class and not chicken embryos. We're just studying chicken embryos. Period. So, the question of what's in our computational image doesn't arise.

There are a bunch of other questions. For example, what does it mean to say that we can "empirically" decide whether something is in our computational image when this appears to be an abstract non-empirical subject? In Haugeland's article that I mentioned yesterday, it's somehow assumed that the question of how to empirically decide whether something is playing chess or understanding Chinese is a meaningful question. And that's not obvious. Then there are other questions which one could raise.

Let's assume that all those questions are answered somehow, and that's a big assumption. I don't think they're easy to answer, barring stipulations, but let's assume they're answered in some fashion.

Incidentally, a comment came up yesterday suggesting that, if you pick the Turing machine criterion for 'in our computational image', then that's behaviourist, and that's not correct. There's nothing in the Turing test that suggests any hint of behaviorism. Behaviorism is a set of ideas about how cognitive states or capacities or whatever develop, and different varieties of behaviorism have their own stipulations as to how that happens – operant conditioning, hidden variables, associationism, etc. Each variety of behaviorism is defined by its *a priori* stipulations as to what modes of changing state are allowed to the organism, and there are extreme versions of behaviorism like Skinner's and Quine's which say you're not

allowed to attribute any structure to the organism. There are versions that are laxer on what you're allowed to attribute to the organism. But these are all theories of acquisitions of dispositions or capacities or something or the other, and there's nothing in the Turing test about that.

Coming back, let's suppose we work all this stuff out, then we have these two approaches – the naturalistic theory (which could be a computer model of the mind), or it could be non-naturalistic (the mental equals the computational approach), and these two approaches are going to be quite different. Of course, we can't really state the differences until the naturalistic one is really worked out – until the gaps are filled – but we're granting now that they are filled.

The two approaches are very different, and you can see it very simply. Suppose we have two theories of rhyme – T1 and T2 – and suppose that our understanding of human cells is such that T1 is realizable in cells and T2 is not. Now, these two approaches would differ entirely on what they conclude on the basis of such a discovery. The naturalistic approach would pick T1 and reject T2, because we're trying to study humans. If we had two accounts of the properties of humans but only one of them is consistent with other accounts of the properties of humans, we would select the consistent one.

The non-naturalistic approach would in principle reject that evidence, because it's a non-biological science and so it doesn't matter whether the mental is represented in cells or silicon chips or whatever. Therefore, the fact that one of them is realizable in cells and the other isn't will just be irrelevant from the point of view of the computer model of the mind as generally conceived. That's independent, incidentally, of whether you take the criterion of being in our computational image to be the Turing test or something else, but I'll stick to the Turing test version because that's the only one that's relatively clear, and that's the one almost anyone accepts anyway.

Anyhow, that doesn't say that the non-naturalistic approach is wrong, it just says that the two approaches are strikingly different. Take the example of ERPs that I discussed yesterday. If one theory of types of classifications of deviant utterances correlates with ERPs and another theory doesn't correlate, and, furthermore, if we had a theory of electrical activity of the brain so that these ERPs weren't just crazy numbers but actually arose from something principled, then in that case the naturalistic approach would select the theory that correlates with ERPs, and the non-naturalistic approach wouldn't care. And the same is true about information from language acquisition, aphasia, other languages, etc. So, across the board, there's

going to be a sharp difference between the naturalist and the non-naturalist approaches.

Well, what kind of reaction should one have to this? With regard to the naturalistic approach, in my opinion, it carries no burden of justification. If somebody wants to study the phenomena that fall under the mental the way everything else in the natural world is studied, that's self-justifying, and maybe one can argue this, but that's my assumption. There's no burden of proof to be borne by the naturalistic approach. There are only questions of whether you're doing it right, or about truth, or something, and then there are the standard methodological questions that arise throughout the natural sciences, but they don't bother us here.

What about the proper reaction to the non-naturalistic approach? There is a range of possible responses. One response is the one given by advocates of the naturalistic approach. I'm talking about an interpretation of the non-naturalistic approach – the mental equals the computational, not a biological science, interested not in humans but in things that are in the computational image of humans. A naturalist (the person who says we should study the mental the way we study everything else in the world) responding to that would say that it's not worth looking at. Unlike the naturalistic approach, it carries a burden of justification. If you say that you're going to study a new problem in chemistry the way every other problem in chemistry is studied, you don't require any justification. But if you say that in this new area of chemistry, you're not going to care about physical realization, then you have a burden of justification. Maybe you can't meet it, maybe you can, but you have a burden at least. But a justification is not given as far as I can see. For example, nobody has given a justification for the decision not to use information about cells to choose between T1 and T2, or not to use information about electrical activity of the brain to choose between two syntactic theories, or not to use evidence from child acquisition to choose between two theories of the mental. As far as I know, nobody has given a justification for that, or, interestingly, even attempted to give a justification.

You can see the reason why people are led to the non-naturalistic approach. It's an exegesis, but it seems to me the reason if you look back at the history of functionalism (as it is sometimes called) is that there was a feeling of uneasiness about the enormous gap between the phenomena that fell under the mental and what was known about the brain. And there were various attempts to relieve that feeling of uneasiness. One approach to relieve that feeling was sort of functionalist – 'It's software, so it's robust things like computers, no big problem'. I think that's a mistake all along the line, for the reasons I mentioned – it's taking a manageable

problem and trying to make it seem less mysterious by using an unmanageable analogue, which raise all kinds of new questions that arise for artifacts but not natural things. As I said, that's like coming out of the frying pan and into the fire. But that's the only motivation I know, if there is any other justification, I haven't seen it in the literature. From a naturalist's point of view, there's a burden of justification to be given. It hasn't been given. It hasn't even been attempted. Therefore, there's no point looking at it any further, just forget it. If any further problems arise in developing the non-naturalist approach (like Searle's Chinese room problem or other ones), the naturalist just doesn't care, because it wasn't worth making the step in the first place. And any problems that arise in making that step could be of intellectual interest but they are of no empirical interest – it has no bearing on the world. It would be like puzzle-solving, which is sometimes interesting.

That's one kind of reaction, and I stress, again, that the naturalistic approach doesn't deny that the brain may be a computer executing software. It accepts the possibility of (and, in fact, most versions accept the truth of) the computer model of the mind, but not the one that's divorced in principle from biology – from the natural world, in other words– and that is studying things in our computational image but not us.

There is another approach which says that this is all meaningless. That's what I would take Wittgenstein (the later one) to be saying. He talks a lot about these topics, and the point he makes is that words like 'think', or any of the terms that involve the mental, are words of our language, and these words are just tools which are used the way they're used. Period. There's nothing else to say about them. He makes a descriptive claim about the word 'think', which I think is accurate – the word 'think' is applied to persons – *persons think* – just like *birds fly*. The question whether something that's not a person thinks is as meaningless as the question whether something that's not a bird flies. It's not the example he gives, but the analogy would be – if someone asks whether it is an objective truth that airplanes fly, a Wittgensteinian would say that that's not a question. The way the tool is used, birds *fly* (and maybe things enough like birds), but not machines – you can decide if machines fly or they don't fly, but that's just changing the use of the tool. You can say that a high jumper flies or you can say that he doesn't fly, but that's just deciding to accept a certain metaphor, which is a matter of decision and not of fact. So, the Wittgensteinian approach says that 'think' is a term that applies to persons or, he says, maybe dolls and spirits (things that are enough like persons), and he's not going to require there to be any sharp criterion for that, of course, because this is just how the tool is used, and the tool isn't used very precisely. But machines don't think any more than they fly or swim or something. The question whether a machine thinks is

as meaningless as whether an airplane flies or a submarine swims. The issue is meaningless for Wittgensteinians. That's another approach.

Notice that that's also a total rejection of the computational theory of mind, but on different grounds. It's not inconsistent with the naturalist's objection. In fact, in my view, both objections – the naturalistic one and the Wittgensteinian one – are correct, but they are different objections, and they both end up with saying, 'Forget about it, it's not worth answering the question.'

A third kind of approach says that it's meaningless but important, and that's not as paradoxical as it sounds. That's Turing's approach, the founder of the field. Recall that what Turing says is that (he's possibly borrowing from Wittgenstein) the question whether machines think may be "too meaningless to deserve discussion", but it's an important question, and then he proposes the Turing test as a 'persuasive definition' – an attempt to convince you to change your usage. He says that we ought to change our usage so that we use 'think' and all the rest of the associated words in such a way as to accord with the Turing test. And he predicts that by the year 2000 educated opinion will have sufficiently changed and usage will have sufficiently changed so that people will find no difficulty in speaking of machines thinking, although today in 1950 it's meaningless (like airplanes flying in 1900's). So that's a view which says that it's meaningless but important, and not a contradictory view – he's saying we ought to talk this way and not some other way. So that's the third reaction.

The fourth reaction says that the non-biological computer model of the mind is exactly right. This view is held by almost everyone who's interested in these questions, that it's exactly right and we have to proceed to work out the problems. Then we get to questions that, for example, Haugeland raised – how do we empirically decide that a given strange object (usually a computer of some kind) is playing chess or understanding English? Having accepted Turing's persuasive definition, he says the answer will be – if it fools the skeptic (if it passes some version of the Turing test). And then he proceeds to give what he describes as an empirical argument – that you can do it for playing chess but you can't do it for understanding English. That's an empirical claim. Well, that's the path you pursue if you think that all this is exactly correct. I want to distinguish between two varieties of this – the approach being legitimate and telling the truth, and the approach being legitimate and being wrong.

The approach is a perfectly legitimate idea, and, furthermore, for some people it's an empirical truth that this is the approach to take. These things are very hard to

put sharply because this non-naturalistic approach is in some hazy area between mathematics and the empirical, so you don't know exactly what is meant when people say they're making empirical claims. It's hard to see how you can be making empirical claims in what amounts to a branch of mathematics, where you sort of stipulate certain things and you don't care about its relation to the natural world. Anyhow, we're assuming all of that to have been settled somehow. Then you could assume that the approach is legitimate and true, and then we have to answer all these empirical questions about machines playing chess, etc. And that's the assumption that holds in most of AI (which is more or less defined by this), most of cognitive science, large parts of philosophy of mind, etc. Though not everyone holds this view. Hilary Putnam, who was one of the originators of this point of view, has recently rejected it totally, not on the grounds I'm talking about, on different grounds.

There's another possible approach, and that's John Searle's. I'm not certain I understand what he's exactly saying but I think he's saying this – it's appropriate and legitimate, but empirically false. And he gives an argument that says it's false. He couldn't be giving an argument that says it's false of he thought it was meaningless – then he wouldn't have bothered. If you take either of the first two views, you don't give any arguments, because it's not worth pursuing so who cares.

Searle's Chinese room argument has had a huge impact on the field. There's an enormous literature trying to respond to it, deal with it, and so on, so it has set off a major current in the field. And, at least as I read those contributions, they all seem to be assuming that the approach is legitimate – that it makes sense to identify the mental with the computational. And then Searle says, 'it's wrong', and almost everyone else says 'it's right', and then the debate rages back and forth. So, Searle appears to be saying that the computational theory of the mind is legitimate – even important, perhaps – but we empirically decide that computers can't think, i.e., algorithms don't have mental properties.

Notice that for a Wittgensteinian, you don't empirically decide that. For them it's just a fact that algorithms don't think because that's the way the tool 'think' is used. Similarly, brains don't think. Brains don't understand Chinese. I understand English, but my brain doesn't understand English, although understanding English is going on in my brain. And the reason for that (for the Wittgensteinian) is that the locution 'understand English' applies to persons not brains. Therefore, though I understand English and I'm doing it by virtue of my brain doing something, my brain doesn't understand English, and, certainly, whatever procedure my brain is carrying out doesn't understand English. It is not an empirical statement, it's just a comment

on what words mean. But for Searle it's an empirical statement – he thinks he's given a proof that brains do understand English but algorithms don't.

Let me repeat his argument which has had an enormous effect. Searle just assumes the Turing test version of being in the same computational image (that's so common that that's fair enough). So, his story is that you've got the Chinese room over here and JS inside it. And the questioner is feeding in questions in Chinese, and the interface is turning those Chinese symbols into a sequence of flashing lights (or whatever the code is). And JS is sitting inside there looking at the sequence of flashing lights.

Searle is supposed to give a reductio argument. He says, suppose that understanding Chinese is done by an algorithm – call it Algorithm A. That's the hypothesis, and he wants to reject it. He says, here is JS and he has in front of him Algorithm A. JS has the capacity to understand its instructions. JS has an arbitrary amount of time and paper and so on. This signal comes in and is translated into a series of flashing lights. The questioner was asking some question in Chinese, like 'what's the weather outside?' or something. JS runs through Algorithm A that by hypothesis understands Chinese. Then at the end JS types out something in some kind of code, and the interface interprets that as an expression of Chinese, and the questioner outside says, 'Yeah, it's the appropriate response, therefore, so far, you've passed the Turing test'. And we're assuming that for every question that comes in, you give an appropriate answer as well as a human would do, in fact, indistinguishable from what a human would do – that's the hypothesis. But, Searle argues, JS doesn't understand Chinese and there hasn't been understanding of Chinese going on. In fact, we could even make a stronger statement – inside that room JS didn't even know that what was going on was translating Chinese, all JS knew is that there are lights flashing and he's supposed to carry out some mechanical manipulations. In fact, putting JS in the room is basically only for dramatic effect. We could put in the room Wong, who does understand Chinese, and he could be doing the same thing, and there still would not be understanding of Chinese going on, because Wong wouldn't have the foggiest idea that this input-output relation has anything to do with Chinese – all he's getting is lights flashing in front of him and he's doing certain mechanical instructions and pushing some buttons. Therefore, the fact that Wong understands Chinese is irrelevant, because there's still no understanding of Chinese going on. Therefore, Searle concludes that we've rejected the hypothesis – that there can't be an algorithm for understanding Chinese – that the brain can't be using an algorithm when it understands Chinese.

We can actually simplify this story. We don't need to take something exotic as understanding Chinese. The same issue arises if you take something quite simple.

Take, say, recognizing a straight line. Suppose somebody proposes that what the visual cortex does in identifying something as a straight line and not a curve is to carry out a certain algorithm. There are such theories of visual perception, and by the same argument you could show the theories are wrong, because you could have JS inside here getting the coding of the straight line or the curve, which is just some series of lights, and carrying out the algorithm and giving the proper answer, but not having the foggiest idea what he was doing –there would be no *seeing* straight lines. And by the same argument you could show that the brain doesn't carry out an algorithm to identify straight lines. In fact, you can push it to the point where it becomes close to paradoxical, though not a literal paradox.

Searle, remember, doesn't think that this argument is strong enough to eliminate the possibility that people ever use algorithms. He thinks that there's a well-defined class of cases where they do use algorithms, namely, those where they have accessibility to consciousness. I don't think you can make any sense of that, but let's grant it. So, there's a class of cases where the brain is really carrying out an algorithm, and that's where the person could introspect and say, 'Yes, that's what I'm doing'.

So, let's take such a case – addition. Suppose JS is carrying out a particular algorithm for addition (like one of the methods for addition taught in school) and he knows that he's doing it. In this case, Searle agrees that the brain is executing an algorithm, but notice that by the same argument, it can't be right, because if you put JS in the room, all he is doing is pushing some buttons when a series of lights are flashed. So, in fact, he will have given the correct answer, and he will have used that algorithm, but he doesn't even know that he's doing arithmetic, all he knows is he's following some input-output instructions. By the same argument we seem to have shown that his brain is not executing the algorithm that Searle agrees it is executing. This isn't quite a paradox because you could reformulate it so that you get out of it, but we're pretty close to one.

Searle's conclusion from all of this is that we have to reject as false (but not as meaningless) the meaningful proposition that the mental is the computational (except in certain cases where you have access to consciousness). So, that the algorithm can't think is a non-trivial fact, which is why he had to give an argument for it. It's not a trivial fact like it is for Wittgenstein (for whom it's just like 'submarines can't swim'). Then comes a huge debate about whether there's some

way around Searle's argument. Curiously, in this whole debate, no one ever studies the simple cases like adding numbers or recognizing straight lines, though they raise all the same questions. So, the whole discussion is about understanding Chinese, which is so far up in the air that you don't even know what you're talking about. Searle's conclusion is kind of like a naturalistic conclusion, but not quite. His conclusion is that since he thinks he has shown that except for the conscious cases the brain isn't executing algorithms, so, crucially, he would reject (and that's part of his main point) anything having to do with I-language, because that would be executing algorithms (although in a naturalistic model).

And Searle ends up in a position that everyone is unhappy with – that it's just some mysterious property of the messy object up here that it thinks. Well, that sounds like La Mettrie, but not quite. It's more like this old story about the brain secreting thought – it's just something about biological mechanisms that gets us to think, though not executing algorithms. Then, of course, the naturalist would ask why not have that property. And there's no answer to that.

The arguments are sophisticated and complex, but I don't see much in it. For example, a philosopher named Stephen Stich has argued that something has to be wrong with Searle's approach, because if you take the brain of somebody who's thinking (or for that matter adding numbers), and since according to Searle this is just a matter of some mystical property of cells, suppose we take one cell in his brain and replace it by a chip that has exactly the properties of that cell. Is he still thinking? You get on down to a slippery slope argument. Presumably, he's still thinking if you change one cell. Suppose you've changed n cells and he's still thinking, is that next cell going to make him stop thinking? Obviously, if there's no point at which he stops thinking, you're going to end up with silicon chips (a computer, in other words). But it's doing the same thing all along. So, there has to be something wrong. And Searle has his answers to those, and so it goes up and back.

I'm not going to say anything about it because it's not worth running through the debate, in my opinion. It's kind of amusing but it's beside the point. It's beside the point because the questions weren't meaningful in the first place. They weren't meaningful for Wittgenstein's reasons, and they're not worth looking at for naturalistic reasons. That's two different things. If that's correct then it doesn't matter much what happened in the debate, and it doesn't matter whether what I just mentioned is a real paradox or only close to a paradox. In fact, nothing matters, because it wasn't worth looking at. It's about on a par with the debate 90 years ago as to whether airplanes can fly. Somebody might've come along and said, 'Yes, if they're enough in the mechanical image of eagles' and had then proposed the

equivalent of the Turing test to fool people into thinking they're eagles or something. That wouldn't be a sensible discussion, and, as far as I can see, this is no more sensible.

I think the reason it's pursued is another example of the temptations of non-naturalism and the reasons for resisting those temptations. If we simply resist them, the questions don't arise. I think we should resist them, and I don't think they're fundamentally problems if we resist them. If we simply go back to the naturalistic approach to mind and we tentatively propose (maybe correctly, maybe wrongly) that there's a computer model of the mind, namely, part of the architecture and the structure of the brain in fact is a computer which in fact is executing such-and-such algorithms in some cases (these being factual statements about the brain on a par with other factual claims about properties of physical objects), we then investigate them by the normal methods of science. No special conceptual problems arise, and that's it. None of these questions come up, and we don't seem to have missed anything, and we have no burden of proof to bear, and we can, if we like, appeal further to the Wittgensteinian argument that there doesn't even seem to be anything to talk about.

Well, that's my opinion about cognitive science, and, again, remember, I'm talking about the reflective parts of it, not the parts where somebody works out the algorithm for recognizing a straight line, which is fine, although cognitive scientists are often schizophrenic about this issue, I think much in the way that physicists were schizophrenic about the molecular theory of matter 100 years ago – the methodological side of their brain says 'We can't do what we're doing' and the scientific side of the brain says 'I'm going to do it anyway.' You very commonly find that.

Well, superficially, at least, the departure from naturalism in the contemporary cognitive sciences is kind of reminiscent of the Cartesian tests for the existence of other minds. So, for example, if you go back to the Cartesian era (around the 17th century), there were people called the minor Cartesians – people who tried to carry out the Cartesian program but didn't make a huge impact, people like de Cordemoy. He set up elaborate thought experiments (in those days everything was a thought experiment, including, probably, what Galileo was doing), which would be used to determine, as he puts it, whether another creature that looks like me has a mind like mine. Crucially, 'that looks like me', because he's tacitly assuming the Wittgensteinian test without the sophistication that goes behind it – if it doesn't have a face, then the question of it's having a mind doesn't even arise.

Remember, the task was this – I know I have a mind because of the Cartesian arguments, and the problem is what's called 'the problem of other minds' – how do we know anything else has a mind? How do I know you have a mind (but a table doesn't), since I can only introspect about me and not you? So, de Cordemoy sets up a battery of tests, and then he says that if this other creature passed all of these tests, it would only be reasonable to assume that it has a mind like mine. That's the argument.

That sounds like the Turing test, but it's not, and the reason it's not is because it's entirely naturalistic. His tests are like the litmus tests for acidity – they're tests to determine whether some object has some real property. Remember, the science that he had behind him (which was wrong science, but science) is that there is a property being a mind which is rooted in a particular substance, and some things have it and some things don't, and he wanted a bunch of tests to determine whether some thing has it. That's completely naturalistic. Wrong-headed, as we know, but naturalistic. That's not true of this approach where everything is crucially non-naturalistic and insists upon it. Though the tests sound similar, they're radically different conceptually.

The same is true of simulation. Actually, there's quite an interesting paper on this by a British psychologist named John Marshall that came out in some festschrift a while back. It's well known that people in the 17th and 18th centuries were utterly fascinated by automata. They had these guys creating very complex automata that would do all sorts of crazy things, and that naturally raised the question whether people are just automata – just as today with computers it raises the same question. Marshall goes through the whole record. The most famous of these guys was somebody named Jacques de Vaucanson, who was building automata that were the marvels of Europe. Marshall goes through the background and he shows quite convincingly that people like de Vaucanson were trying to carry out simulations in the modern scientific sense, not in the computer science sense. One of de Vaucanson's examples was a duck digesting food, and he tried to model the process of duck digesting food in an automaton, and it was so realistic that everybody was amazed. But crucially, he was not trying to say that the duck *is* digesting food.

A replica of Vaucanson's mechanical duck created by Frédéric Vidoni for the now defunct Grenoble Automata Museum. Credit: mus-col.com

Think what he could've been doing if he'd been doing the analogue of the computer model of the mind. He could've said – I'm interested in the digestion of food by ducks as a non-biological subject. I abstract away in principle from ducks. I'm just interested in digestion of food by ducks abstractly conceived. I want to consider all things in the mechanical image of a duck, and I'll say all of those things are digesting food if some criterion is satisfied (some analogue of the Turing test – that they fool people into thinking they're ducks or something).

That would be an analogue to the contemporary cognitive science and philosophy of mind. But he didn't do that. In fact, he purposely made the duck transparent so that people could see that it's not a duck. He also wanted them to be able to see what was going on in his model. And, in fact, what he was doing was constructing a model very much in the normal manner of the natural sciences –

here's a model which he thinks gets down to the core of what digestion is, and he's going to demonstrate these core properties by having you look at the model.

Well, that's simulation in the sense in which it's done in the natural sciences, and it's very different from the kind of simulation (if you like) that's done in chess-playing programs – the drosophila of the cognitive sciences, as Herb Simon put it – where the trick is basically to achieve some criterion (in the usual case to fool people into thinking they're a person). That's a completely different form of simulation. For the 17th and 18th century people, the chess-playing programs would've been of no interest whatsoever, because they're teaching you nothing about how people think. Zero. They're designed on different principles, therefore they're of absolutely no interest. They would've taken exactly the view that I think everybody ought to take to the chess-playing programs today, and certainly any natural scientist should, and would. They're just a waste of time, they're not teaching you anything about people in principle or about anything else in the natural world in principle. Rather de Vaucanson and others were interested in simulation in the way it is done all the time in the natural sciences – what I call contingent or exploratory abstraction – you throw away some properties and look at other properties because you figure that way you can learn something about the real object. That's like the computer model of the mind taken naturalistically.

In my opinion, there's been a very serious regression since the 17th century in this regard. The cognitive sciences picked up many old topics that were forgotten, and that is good, but, in my opinion, they're approaching them much more irrationally than they were approached in the 17th and 18th centuries, and this comparison strikes as a good example of that.

The computer model of the mind is so completely taken for granted in the philosophical literature that it's led to very serious misunderstandings about what other people are doing, because it appears to be almost inconceivable in the literature that anybody could be taking a naturalistic approach. So, what you get is efforts to interpret work (like mine, for example) which takes a naturalistic approach to the mind and people get really confused and they see huge paradoxes because they take it for granted that I must be talking about the computer model of the mind in their sense because that's the only possible idea one could have.

There are some good examples of this in various publications by two philosophers named Michael Devitt and Kim Sterelny, who've written a lot about these topics. I'm a target in this case. They develop the concept of competence in English – competence in English is explained in terms of the computer model of the

mind (like a Martian with silicon chips could have competence in English), and then they say that they find it bewildering that linguists (meaning me) should confuse a theory of I-language with a theory of competence that includes Martians competent in English as well as humans, and they take it for granted that that's what we must be studying.

We must be studying a notion of competence which is not parochial, it doesn't just deal with humans but deals with anything in the computational image of humans, and it's non-biological. And since that's what we must be studying, competence in English must be that. And the Martians have competence in English and plainly a theory of I-language isn't about that, because maybe the Martian is doing it some totally different way and not in terms of the principles of UG.

My theory of rhyme or some theory of sentence structure or semantics or whatever is purposely a very parochial theory, because it's supposed to be – it is a theory about humans and their organization. And competence in English is just defined as the property of having a certain state of the language faculty (which is this range of I-languages, because English isn't a thing, of course, but an indeterminate collection of things), but that seems to them an unintelligible notion. Therefore, they find it bewildering that people should give a theory of I-language and confuse it with the topic of inquiry, namely, competence which includes Martians.

Well, I also find it bewildering to confuse a theory of I-language with a theory of competence in that sense, and the reason is I find it even more bewildering to imagine a theory of competence is their sense. What's a theory of competence in English that includes Martians who do it by some method other than I-language? I haven't the foggiest idea what that even means. It seems to me a totally bewildering idea. It's like a theory of digestion that doesn't just deal with ducks and people and so on but with arbitrary creatures who are like ducks by some criterion we set up. That's totally bewildering to me, as any non-naturalist approach is, but there's nothing bewildering about talking about a theory of I-language. There would be something bewildering about extending it to a theory of competence in the sense that they take for granted. That seems to me unintelligible.

That's the standard line of debate that goes on in the philosophical literature. It's like an impasse. People aren't even talking to each other because of this sharp break over the legitimacy or the meaningfulness of the non-naturalistic dualistic approach to the mind that's developed through functionalism and into the computer model of the mind and so on.

Lecture #8

FOR THOSE OF you who are familiar with Searle's work, which I only alluded to briefly, let me clarify certain points of agreement and a few points of difference. I agree with him when he says anything could be a computer. When he goes on to say, however, that his Chinese Room argument shows that the mind is not in any serious sense a computer, there I disagree. His point is that since anything can be a computer, the brain is a computer (just as a pile of sand on the beach is), but there's no serious sense in which it is. And he argues that his thought experiment demonstrates that the brain doesn't calculate algorithms, therefore, it's not a computer in any sense other than in which a collection of stones is a computer. But there, I think, that's because he insists on the non-naturalist approach.

From a naturalistic point of view, whether the brain is a computer with certain hardware and software executing a certain algorithm is simply a question of fact, and the answer is – it is if that's what the best theory tells you. If the best theory of the brain at the appropriate level considers it to be a computational device executing a certain algorithm, then that's just a fact, just as if the best theory tells you something about curved spacetime or whatever. There's no other criterion for being an acceptable statement of fact in the post-Newtonian world. And I think there's good evidence that in fact the brain does – in particular, in language, aspects of vision, etc. Also, another point of difference is that I don't see any reason to accept this criterion of access to consciousness even if it had to be made coherent.

We now turn away from the cognitive sciences and the modern cognitive revolution to another hotly debated issue that seems to me, again, to reflect a lingering non-naturalistic dualism. This is the question whether mentalistic talk and mental entities will eventually lose their place in our attempts to describe and explain the world – I'm quoting Tyler Burge from a quite interesting review of the history of the contemporary theory of mind in the centennial issue of *The Philosophical Review* a couple of weeks ago, which actually puts all these things in place in a very coherent and lucid way for anyone interested in seeing how philosophers have looked at these questions for 50 years or so. He raises the question whether mentalistic talk and mental entities lose their place in our attempts to describe and explain the world or are they going to survive the transition to natural science. That's kind of an odd question to ask, I think, and you can see that by changing 'mentalistic talk and mental entities' to 'physicalist talk and physical entities.'

Suppose someone were to ask whether physicalist talk and physical entities would survive the transition to natural science. That's a question that everybody knows the answer to – 'No, except by the most colossal accident'. The concepts of natural language are simply not appropriately designed for the study of the natural world, at least the so-called physicalist ones – the ones that refer to things like bodily motion, etc. To take the example I mentioned a couple of times, if a physicist is trying to decide whether a pile of sand is gas, liquid, solid, or some other state of matter not yet properly characterized, the physicist does not look at the way ordinary people use the word 'liquid' or 'gas'. It's absolutely of no interest. There is no such thing in philosophy or in the natural sciences as 'liquid realism' or 'falling realism', meaning the belief that notions like 'liquid' in the ordinary language sense (in the sense of folk physics) have to do with real things, things of the natural world. The questions of reality and of natural kinds only arise within the natural sciences, they don't arise within common-sense discourse. So, there's no liquid realism, momentum realism, falling realism, etc. Nobody cares in the natural sciences whether the concepts they develop happen to match some of those of ordinary language, and nobody expects that they will, short of miracles.

Nevertheless, there is a debate about something called 'intentional realism', and that's what Burge is talking about. The leading exponent of it is Jerry Fodor. Intentional realism gives a positive answer to the question that Burge raises – it says that mentalistic talk and mental entities will survive the transition to natural science, and must, that is, the true psychology (the true theory of mind) will meet the conditions of the notions of folk psychology. The assumption of intentional realism is that the concepts of folk psychology (beliefs, desires, referring, etc) have so robust a status that they're going to survive the transition to science, unlike notions like moving, falling, liquid, etc., which we know aren't going to survive the transition. So, nobody expects folk physics to be carried over into the natural sciences, but there's a huge debate as to whether the concepts of folk psychology give us precisely the right framework for the study of the mind. Burge suggests that mental entities and their talk won't survive the transition to natural science, and Fodor argues they will, and this is the debate over intentional realism. But the whole discussion is curious – nobody assumes that in the case of the physical, so why should anybody raise the question in the case of the mental? Again, it seems to me a relic of the kind of epistemological dualism which gives a sort of priority to the mental that the physical doesn't have.

Let me turn to a totally different class of questions that have been coming up now and then in the discussion, and now I want to finally review them – the class of questions that are raised with regard to such notions as I-language and

having/knowing an I-language. So, these are the two notions we're going to look at – I-language, which is this internalist conception of language, is a state of the brain; and having/knowing/speaking an I-language is being in that state. So, an I-language is a state of the language faculty, having an I-language is being in that state, and using an I-language is having your performance systems access and make use of information that's stored in that state. That's the internalist view of language; I'd say a naturalistic view.

Now, there are a lot of problems that have been raised with regard to this. It's considered to have been demonstrated to be wrong. It's not only not accepted, but the issue doesn't even arise, and there are some arguments for that. Some of the best arguments are given by a young philosopher named Alexander George, who has written about this a lot and happens to know linguistics well too. There are other arguments by Michael Dummett, a very distinguished Oxford philosopher. The arguments go kind of like this.

Let's take some North American named John. John knows English. Luigi knows Italian. John has beliefs about English and Luigi has beliefs about Italian. John speaks the same language as Mary, but Luigi doesn't speak the same language as Mary. Languages change and evolve and so on. I seem to be referring here to language all the time, but I couldn't have been referring to states of the brain. So, take, say, John's beliefs about language. Certainly, John has beliefs about his language, but he doesn't have any relevant beliefs about the state of his brain, that's for sure. We seem then to have shown that there's some notion of language independent of the states of the brain. The notion of language which has entered into discourse of this kind can't be a state of the brain. States of the brain don't change and evolve. John doesn't have the same state of the brain as Mary, but he speaks the same language as Mary. John has beliefs about his language, and he doesn't have any relevant beliefs about the state of his brain. Therefore, it follows that language is different from states of the brain (i.e., I-language). Furthermore, 'having a language' must be this notion. And that's relevant, since this concept of language is the one that can be shared. John and Mary can't share the states of their brains – that's a particular arrangement of molecules – but they can share a language, and therefore, language can't be a state of the brain.

So, the conclusion is that the basic notion of language can't be I-language. It must be what we might call 'externalist' – something outside the head – and then this carries over to everything else – to reference, to meaning, to perceptual content, etc. All must be externalist. It's argued that psychology itself has to be externalist. It can't study what's in people's heads because it can't capture notions like these if

it does, and these are the notions that matter. There is some entity called (say) English that's outside my head and that I share with (say) John Smith, and each of us stands in a 'cognitive relation' to it, but it's outside us, it's an abstract object – what's sometimes called a Platonic object – and we stand in a cognitive relation to it since the object is shared, and we differ in behaviour.

Suppose John Smith comes from Oxford. I come from Boston, so we differ a lot in the way we talk, hence we don't share behaviour, but we both stand in a cognitive relation to this thing called English. It therefore follows that we have only partial or erroneous knowledge of English, as Dummett puts it, and the task of the linguist is to explicate the cognitive relation that holds between John and English – that abstract object – which could be either a Platonic object (like, say, a number), a community property, a social practice – something that exists independently of any particular speaker. So, whether somebody has this language or not, it exists, just as whether anybody is thinking of the number 87 or not, it exists in Platonist mathematics. English exists independently of any speaker, and each speaker, typically, has a partial and a partially erroneous grasp of the language. So, I don't really know English, I have only a partial grasp of it, and that follows from the fact that I share it with the people who I talk to, like, say, my wife, and, so we talk somewhat differently since we have the same language but different behaviour. Therefore, it follows that each of us has at best a partial and hence partially erroneous knowledge of this thing.

Some people put it a little differently. James Higginbotham (a colleague of mine) has suggested a position he calls 'weak conceptualism'. We must, he argues, accept the idea that there are these abstract objects that we stand in a cognitive relation to, otherwise we can't run through these apparently valid arguments. But he recognizes that English has no properties other than what is the reflection of an I-language. He suggests we think of it as a kind of Platonic shadow – the mapping of the I-language into some abstract domain – so you could then stand in a relation to it and can have beliefs about it. You don't have beliefs about the state of your brain, or if you do, they're not relevant. But you could have beliefs about this abstract image of the state of your brain – and that's what linguistics must be about. That's the general idea.

Notice that this is again a non-naturalistic argument. Crucially, it's claiming that the core notion of language is not naturalist, that it's not part of the natural world, it's an abstract thing – either a Platonic entity or a community practice or a collection of norms which is not a thing of the world. Hence the approach is non-naturalistic. It therefore, as usual, requires that a justification be given. And in this case a

justification has been given, which is different from other examples I cited where no justification is given.

Just to repeat one of those several arguments here. We have beliefs about language. We have no relevant beliefs about the state of the brain. It therefore follows that language is not a state of the brain. Since I know English, I must stand in a cognitive relation to this thing, and since the object is shared but the behaviour differs, we have only a partial or erroneous grasp of the object. In fact, it may be that everybody who speaks English has only a partial or erroneous grasp of English (which is what it is, independently of any English speakers past or present).

The model for most people here is the Platonist view of mathematics. This view of mathematics is not accepted by every philosopher of mathematics, certainly not by working mathematicians, but the Platonist view of mathematics says that our relationship to mathematical entities (numbers or sets) is kind of like visual perception – we perceive a table and we grasp the concept – and among the concepts we can somehow grasp are those that constitute the entities of mathematics. And this is a world of things, and even if we're not grasping them, they're still there. Before humans developed, they were there, and there were certain truths about them (say, the prime number theorem), although nobody was grasping them. And sometimes when we discover the truths about them, it's not an invention (obviously), it's somehow a discovery, because when people work on it they feel like they've discovered something.

Sometimes it gets kind of tricky. For example, it's now known that in the case of set theory, there's no axiomatizable version of set theory – you can't capture the truths about it. Therefore, there are questions like (say) the truth of the continuum hypothesis, or the axiom of choice, or various assertions about sets, etc., since according to the Platonist view, they're either true or false and we have to discover their truth or their falsity the same way we discover facts about the world. Kurt Gödel, for example, one of the great mathematicians of the century who advocated this view very strongly, held that the truths of arithmetic (about, say, numbers) are like the evidence that helps you discover the truths about sets. I don't want to raise any questions about this picture – if it's right, it's right, and if it's wrong, it's wrong – but whatever the validity of it, it doesn't carry over to this domain.

But let's accept the Platonist view for the sake of the discussion. Anyone who accepts a Platonist view assumes there are truths about numbers, and they even agree on what they are, and they largely agree on truths about sets. They know there are some things you can't settle so simply, but they assume that there are truths, and

most of them they can find and agree about. They can also tell us what they're talking about. They can tell us what the numbers are and what the sets are, like, maybe the sets are the things that satisfy some axiomatic version of set theory (say, Zermelo-Fraenkel set theory) plus those others that don't follow from it but are the true version of them, like, if the axiom of choice happens to be true, you have to add that one in, and if false, you don't. They can tell us an awful lot about the objects they're talking about.

In the case of language, however, the Platonic objects that we are supposed to be standing in some cognitive relation to, nobody can tell you anything about them. If I ask what's English, nobody can answer. Is English my pronunciation or of John Smith from Oxford? What's the right syntax of English, mine or somebody else's?

Take lexical items (words). There's a word 'livid' which occurs in exactly one phrase in English, namely, 'livid with rage', and just about everybody who learns English assumes that 'livid with rage' means 'flushed', because that's what people get when they're angry – your face turns red. Maybe 99.9% of the population thinks that 'livid' means 'flushed'. You go to school where they teach you the truth, and you write down 'livid means flushed', and the teacher says 'livid means pale', so 'livid with rage' really means 'pale with rage' or something like that.

Take the word 'disinterested'. Practically everybody uses this word to mean 'uninterested' – if you're not interested in something, then you're disinterested – except for people who believe what they've been told in school, where they tell you that 'disinterested' means 'neutral' – that you just have no opinion of something.

Well, is there a truth about this? Is it true that 'livid' means 'pale' even when everybody thinks it means 'flushed', and 'disinterested' means 'neutral' when just about everybody thinks it means 'uninterested'? If yes, then somebody has to explain what all that means. How do we determine the right meanings or the right pronunciations? Why is my pronunciation, which is some urban dialect, wrong for English, while somebody else's pronunciation is right? In other words, what is the true Platonic object and how do we decide what it is?

In the case of mathematics, the Platonists can answer a lot of these questions. Not all of them, like, they can't tell you whether there exists a set with the property asserted by the axiom of choice – it's a question of the existence of certain kinds of sets, and the Platonists aren't sure whether there exist such sets.

In the case of English, nobody can tell you anything. Take, say, Jerry Katz, who is a leading exponent of this field of philosophy. He argues strongly that English

has to be a Platonic object, and it's real, and all of us have a partial knowledge of it. But if you ask him what it is, he just can't say – it's just that thing out there which we stand in a cognitive relation to. That's quite different from Platonism in mathematics, whatever one thinks of it.

Remember, this is an unusual case of non-naturalism, in that an argument has actually been presented. In fact, the example I talked about is the first case where somebody has actually presented an argument for non-naturalism. So, there is a cogent argument, and we have to ask whether we can explain everything that argument claims to explain in I-language terms, and we also have to ask whether an alternative has really been offered. True, there's an argument, but has an alternative been offered? Has a proposal actually been made? If nobody can tell me the simplest thing about English, the thing that I stand in a cognitive relation to, then what proposal has been made? That seems a fair question. If language is a community property or a social practice, then which community and which practice is English, mine or somebody else's? The questions surely arise, and we have to see what we can say about them.

Now, when you ask people to identify the Platonic objects, there's two kinds of questions you're asking. One question is – can you tell me something about this Platonic object called English that I have a partial and erroneous knowledge of? Then there's a more general question – can you tell me what sort of a thing this Platonic object is even if you can't tell me which one it is that I have a partial relation to?

To the first question, there is a kind of an answer, but it's so absurd that there's no point pursuing it. The answer that's usually given is that English (the real stuff) is what it says in the Oxford English Dictionary and Fowler's English Usage. That's just ridiculous. For one thing, they don't even begin to characterize English. They say nothing about any of the properties of the language that I was talking about. They have nothing about syntax. They don't even know the topic exists. Furthermore, they're just stipulations. Everybody knows they're just stipulations made by various power figures sometime in the past – sometimes crazy stipulations which were totally inconsistent with the nature of the language, made for one or another authoritarian reason. Maybe they are plausible, maybe not, but it's just a power system – something you can decide to adhere to if you like, or not if you don't like, but obviously it has no status in this discussion. That ought to be a triviality. So, we can put away that proposal.

So, we're left with nothing that's worth even looking at about the nature of this particular entity called English, Catalan, Swahili, or whatever. As to the general

issue of what kind of thing is this Platonic entity, there are some proposals. For example, there is a proposal by David Lewis, a philosopher at Princeton. He says that these Platonic entities are collections of norms and conventions, and we spell out those notions in terms of regularities in behaviour, or in terms of principles adopted by the speaker on the basis of certain beliefs about other people for the purpose of facilitating communication. So, for example, I adopt certain norms, principles, etc., because I believe something about other people that I want to facilitate communication with. And in those terms, he says, we can come to understand what this external entity is.

There are a lot of problems with that. For one thing, if 'regularities in behaviour' means things that happen with more than zero probability, then there just are virtually no regularities in behaviour. In linguistic behaviour, about the only regularity in English is that when you walk into a room you may say hello or something, and there are probably ten things like that. But apart from that, there are no regularities. When you try to investigate people's linguistic behaviour, it's just always different. There are minor fluctuations here and there (standard clichés), but there's virtually nothing you can say about regularities in behaviour, at least if regularities mean things that happen in particular situations detectably more than chance. So, that doesn't seem to help very much. As to the idea that people adopt certain norms because of beliefs about others that they've worked out in interest of facilitating communication, that doesn't seem right at all. I mean, nobody adopts these norms, and nobody even thinks about them. If you take a three-year-old kid who has all this stuff in his head, it hardly makes sense to believe the child decided, 'Okay, I have certain beliefs about the way my mother talks, and I'm going to adopt certain norms (norm means putting your fork on the left or wearing a tie to class or something) because that will facilitate communication with my mother. I mean, if anything like that goes on, then it's certainly deeply hidden. There isn't the slightest evidence for anything like that and it's just ridiculous. People don't make decisions about these things, it just grows in their heads, and if it facilitates communication, fine. If it doesn't, too bad. Try to get around it somehow.

So, we're basically left with nothing. We're left with no general account of what Platonic English might be and, even more annoyingly, no specific description of any of its properties – we're told that it exists and it must be the topic of linguistics and not I-language, but nobody can tell us anything about what it is, and they can't tell us anything about how to find out. So, if that's linguistics, we better close the books and go home, because we won't know anything about the things and we'll never find out anything about them, at least according to the Platonists. This is again radically different from Platonism in mathematics, which doesn't end up with the

conclusion of closing the books and going home, in fact there are lots of obvious projects to carry out there. This is different than the other forms of non-naturalism that I discussed, in that at least there's an argument. So, we might try to inspect the argument. But it is the same one, in that it just makes no sense at all, as is typical in non-naturalistic approaches.

Remember the way the argument it looks. Take the strongest version of it. John has beliefs about English. John doesn't have any relevant beliefs about the state of his brain (I-language). Therefore, it follows by simple logic that English is not an I-language. That's a straightforward argument. So, we do have an argument, I want to stress that, it's not that no argument is given. But the question is what this argument amounts to.

Let's take something else. John has beliefs about health. For example, John believes that tomatoes cause cancer (and a lot of people believe that (which is probably true because they're highly carcinogenic), but it doesn't matter for the moment whether the beliefs are true or false). Plenty of people believe they shouldn't eat meat or something like that. So, people do have beliefs about health, like, you should run six miles a day or something. So, John has beliefs about health. Now, does it follow that there is an entity in the world – health – that John stands in a cognitive relation to?

John can have beliefs about style of dress, like, he believes you ought to wear a tie when you go to class or something, then is one of the things in the Platonic world style of dress? John has beliefs about dining, he thinks you shouldn't drink white wine with meat. Is there such a thing as dining that John has beliefs about? Or, say, John has beliefs about national rights, he thinks the Welsh should have independence. Is there such a thing in the world as national rights that John stands in a cognitive relation to? You can make these things up as you like.

Turning to knowledge, suppose John *knows* the construction business like the back of his hand, he makes a lot of money when he works on construction. Is there an entity – the construction business – that John stands in a cognitive relation to? Or, suppose John knows the secret of happiness, like, why is he so blissful all the time. So, is there an entity – the secret of happiness – that he has a cognitive relation to? Well, there are people who answer 'yes' to all these questions, but something's clearly going wrong. There's no reason whatsoever outside the mistaken theory of belief and knowledge and reference to assume that there's an entity called 'health', or the entity 'the construction business', or the entity 'style of dress', and so on. To have beliefs about something is just to believe in particular things which happen to

be in a certain range. It doesn't mean that the thing you have beliefs about exists. That's just a gross misunderstanding of language. And to say that knowledge of the construction business is standing in a certain relation to things is just a mistaken theory of knowledge – it's not what knowledge means.

When you look at the argument, you see that the abstract entities that are established are serving no purpose whatsoever. They play no role in explanation. They don't help us understand anything. Nobody can identify them. The only thing they're doing is allowing an obviously mistaken theory about belief and knowledge to grind its wheels pointlessly. It's kind of amazing to me that 40 years after Wittgenstein this can go on. Everybody has read and studied him, and the whole of his later corpus is an effort to make people understand that this is perfect nonsense, that you shouldn't be misled by language into thinking that if there's a word, it has to refer to something, and if you have a belief about something, there has to be a thing you have a belief about. In fact, the critique goes back much further, but certainly it's an astonishing fact about contemporary philosophy that people who have read Wittgenstein, and are influenced by him, and worship at his shrine continue to make exactly the mistake that he devoted his whole later career to trying to undermine a false picture of language (what he called 'the Augustinian picture') that leads you into absurdities like this. He almost defined philosophy as the act of being led into absurdities by a mistaken picture of language.

Notice that this is all different in Platonic mathematics. Whether it's right or wrong, it doesn't have any of these properties. It's not leading you into absurdities. You might not be happy about the idea that there's a Platonic universe of real sets including sets that satisfy the axiom of choice or not (as the case may be), but you're not led into absurdities by that, and in fact there's a certain plausibility to it – it accounts for things like discovery, it gives you an account of truth, it gives you research projects that you can pursue, and so on.

Here there are no redeeming features at all. It just allows this theory of belief and knowledge to grind along, and these theories are just obviously mistaken. That's not what belief and knowledge are like. They're something else. In fact, having knowledge of something probably is nothing more than being in a certain cognitive state, i.e., having the relevant part of the brain in that state. So, having knowledge of the construction business (we can't say a lot about it) is some state of the cognitive system, and if you can spell out the nature of that state, you'd know what it means to have knowledge of the construction business. And things like having belief about health are perfectly trivial to explain.

Well, if this is correct, there's nothing left of the argument. The argument was leading to Platonist linguistics, that is, to the principle that linguists (unlike every other scientist) have to study an abstract entity that cannot be specified and for which no advice can be given as to how to learn more about it. Instead of being left in that uncomfortable position, we can simply unravel the argument and say that it went wrong at the first step, and analogies to other cases should make that quite clear.

Now, in the study of I-language, you may or may not come along with a concept that has some close relation to the term 'language' in ordinary usage, but you don't care. In the study of psychology, you may or may not come along with a notion that stands in a close relation to the intuitive concept 'belief' or 'desire', but you basically don't care. You know it's not going to be true when you talk about 'falling' or 'liquid', so why should it be true here? If it doesn't, okay, then we're just in the same status as the natural sciences are generally. There's nothing wrong with studying folk psychology. It's a fine topic, just like you can study folk physics or folk biology. In fact, there are fields who do that, like, ethnobotany studies people's beliefs about botanical systems, but nobody confuses it with biology. Those are fine topics, but they don't determine the nature of scientific inquiry.

What about the other arguments? I mentioned one based on belief. It is perfectly true that we use terms like 'English' and 'Italian'. We also use terms like 'household pet' or 'terrestrial mammal' but we don't think that they pick out natural kinds – things out of which nature is constituted, like atoms or elements or particular well-defined species. Nematodes could be a natural kind because the science has advanced enough so that it's picked out something that's sort of real, but household pet or terrestrial mammal certainly isn't a natural kind. It's a perfectly usable term but it doesn't pick out any natural kind. We can say that giraffes have evolved longer necks over time, but that translates into something else, that translates into the perfectly intelligible statement that the distribution of neck lengths ten million years ago was different than the distribution of neck lengths now in animals that have a certain historical connection if you look at the sequence of generations. That's what it means to say that giraffes evolved longer necks. Nobody is confused into thinking there's an entity called giraffe which has changed over time, and there is no reason to be confused about believing that there's an entity called Italian that has changed over time. If we say that Italian has changed over time, we mean that the distribution of traits (ways of speaking and so on) in a certain region used to be something and now it's something else and there's a historical connection, that's it, there's nothing more to say. When you say that the nematode neural system has been worked out, what that means is that there's a certain reasonable abstraction from individual nematodes that has certain properties. That's all it means. Some nematode might

have 299 neurons instead of 300, but nobody is confused about that. Similarly, if you say that English differs from Chinese in the rule of question formation, that's going to translate into something about a certain collection of I-languages and another collection of I-languages and some patterns that they share and some patterns that they don't share.

What about a common shared language? That was one of the other arguments. John and Mary share a language but they each have only partial or erroneous knowledge of it, as you can see from fact that they differ in the way they behave. Here, again, I stress that no sense has been given to the notion of a common shared language, and it is in fact quite clear that you don't need any such a notion. Suppose I say that John and Mary look alike, does it follow that there exists in the world a shape that they share? Obviously not. To say that John and Mary look alike is to say that if you look at the relevant traits, they kind of overlap (more or less, because there's never any answer to how much). To say that John and Mary share the same language is exactly the same. They more or less share certain patterns and traits enough so that they can communicate without much trouble, and, again, there's no answer to how much they must share to be the same language any more than there's any answer to how much they must share to look alike. These are just questions of decision relative to particular interests and concerns, and there will never be an answer because it's not a serious question. It's not that you can't use these concepts in ordinary discourse, like, it might be perfectly reasonable to say John and Mary speak the same language but Luigi speaks a different one, meaning that the shared traits of John and Mary relative to my current interests are overwhelmingly high as compared with the ones they jointly share with Luigi, which are relatively low. Similarly, I can say that John looks like Mary but not like Luigi on the same grounds. That's about all there is to it.

There are all kinds of subcultures, they have their particular practices, particular demands, authority structures, etc., there are different colours on maps, there are oceans that separate things and make it look as if there's some natural break, etc., and there are all kinds of ways of associating oneself with other people in what you can call a community, but there are no such things as communities in the natural world. The world isn't divided into communities, just like the world isn't divided into regions. It is perfectly sensible to say Boston is near New York and far from London, but it doesn't mean there are regions into which the world is divided and Boston and New York fall within one of them and London falls outside that. Nobody is led into those confusions. And exactly for the same reasons, if people more or less speak alike, then there's no reason to think that there are regions of common speech they fall into called languages or shared social practices.

Notice that it's not a matter of vagueness. It's not like the notion of bald where you have no answer to the question of how many pieces of hair it takes. It's not even like the notion of think, where, as Wittgenstein pointed out, there's no precise border around the class of things that we can say think – we say it of persons and we'll say it of things are near enough like a person, but it's a vague boundary. But that's not what's going on here. When I say the world does not have regions or areas, it's not that the areas are vague, it's that they are anything you like relative to your current purposes. Exactly the same is true when you say the world does not have languages – it's anything you like relevant to your current purposes. Maybe that the continuity of systems has been broken by some kind of accident – like a mountain, a conquest, or because nobody ever started speaking them or whatever – but that's irrelevant, obviously. Linguistics can't be the study of all geological and historical accidents. It is the study of something about language, and that's just what's in the head, nothing more. It's not a matter of vagueness, it's a matter of the non-existence of the topic of common language. It's one of the pictures where nothing is to be resurrected from.

The arguments that are given are intelligible, but collapse as soon as you take them apart, unless you want to accept all these ridiculous consequences (that among the things in the world are health, style of dress, the construction business, etc). You can see what's wrong with the arguments, and you can also understand why nobody can tell you a thing about the entities that they're proposing we must study. If somebody told a geographer they must study actual areas, they couldn't tell you anything about them, because they don't exist. And the same is true here. To the extent you can say anything about common languages, it's a reflection of what you say about particular I-languages. And there's no question anybody's ever thought of for which there's even a mildly plausible answer where that can't be stated in terms of the real things, namely, I-languages.

So, it seems that this is a picture that simply must be abandoned. If so, there's a lot of stuff that can be thrown out of the libraries, because virtually all of the literature in these fields resorts to the notion common language, shared practice, etc., and it's very hard to find something that doesn't. Like I said, externalist approaches are taken to have been established, and they aren't even debated much anymore. So, you should think about it and see if it's exactly right, because there's a lot at stake in the study of language, psychology, and so on. And I stress again that no one has ever even tried to give answers to the most obvious questions, like, if there is Platonic English then what is it, and so on.

Another question which has a large literature related to this is – what is linguistics about? There are a lot of papers with similar titles, or papers asking what's the right theory of linguistics, or whether it's the right theory, etc. It's a big topic of debate and there's a range of positions. The topics we've been discussing fall into that. The almost standard view says that what language is about is these abstract entities or shared practices or whatever. It's true that linguists go around studying I-language, but they're just confused, and they should get out of that confusion and start studying the real thing (about which they can tell you nothing and about which you'll never find out anything). That's what they should do.

There are even more extreme views than this. Devitt and Sterelny, in responding to a paper of Fodor about what linguistics is about, take the position that to find out what linguistics is about, we should ask Grandma (any arbitrary person, in other words). So, we should ask Grandma and we should adopt 'Grandma's view'. Well, what's Grandma's view? Grandma has no view at all about how people speak and understand, so we're not allowed to study that topic. We're only allowed to study what Grandma says, and what she probably says is language is about words, and meaning of words, or something like that, so that's what we have to study. Again, this is a very curious position, right or wrong. It's interesting that no one suggests that biologists should ask Grandma's view of what they should study – just linguists.

Again, that's a sort of radical dualism. But there's even a deeper question. There is no question about what biology is about – there's no article saying 'What biology is about' – because it's nonsense. For example, is organic chemistry part of biology? From a certain point of view – Yes. From another point of view – No. Is physical chemistry part of physics or chemistry? Those aren't questions. In the sciences, everyone understands that the professional disciplines don't carve the world up into separate domains, they don't cut nature at the joints, so to speak. They're just conveniences. They're there for university administrators or for grant agencies, but they have no meaning at all. Science is just trying to find out about the world, and if things cluster in particular ways at a particular time, okay. If they start clustering some other way, okay. There is no topic that asks what biology is about, but there is a topic asking what linguistics is about, and people argue up and back about it, and they don't tell us why. Incidentally, some people, I should say, do believe that the professional disciplines like chemistry and physics really do reflect ontological differences about the world – Jerry Katz, for example – not only he believes it, but he seems to regard it as obvious. It's pretty hard to make any sense out of that. I can't imagine any chemist or physicist believing that.

There's another argument for common public language which is pretty standard. The idea is that there's a public object out there – English – that I stand in relation to and Mary stands in relation to. The argument is that the notion of common public language is required in order to explain something that happens, viz., communication. Surely, communication happens, therefore, if we don't have some kind of shared language, how can there be communication? This shared language is going to have to have shared meanings and shared reference.

Curiously, nobody ever suggests that the shared language has to have shared pronunciation, and it's again that same dichotomy. When you talk about the sound side of language, it's somehow much less tempting to fall into total irrationality than when you talk about the meaning side. I think that's the residue of the same dualism again.

Anyhow, the argument is that if you don't have a shared language with shared meanings and shared references, how can you communicate. Fodor and Lepore, in their book about meaning holism argue, as follows. They argue that linguists can adopt an I-language perspective only "at the cost of denying that the basic function of natural language is to mediate communication among its speakers." So, if we adopt an I-language perspective, as I recommend, then we have to deny that the basic function of natural language is to mediate communication. That assumes, notice, that successful communication between Peter and Mary implies the existence of a common language that they share. But that's no more obvious than the fact that physical resemblance between John and Mary implies the existence of a common form that they share. It's obviously false in the latter case, and it's not obvious why it's true in the first case.

What about the cost of denying that the basic function of natural language is to mediate communication? That statement, which is not uncommon, is quite unclear. What is the absolute notion or the basic function of any biological system? Systems don't have functions. They do things, but they do lots of things, and it's very hard to see any sense in the idea that this is its basic function. Sometimes you can make a comment about that and sometimes you can't, but it's not a very clear idea. And even if we were to agree that language has a basic function, why is it to mediate communication? After all, it's just one of the many things done with language. Why isn't it to think, or to express your thoughts, or to tell jokes, to have social relations with people where you don't even care whether you communicate, etc.? Those questions will have to be answered, but there's no point pursuing that course because there's no problem at all in accounting for the possibility of

successful communication in strictly internalist terms, insofar as an account is possible at all.

Insofar as there's an account of communication, it presumably works something like this. So, let's take John and Mary. John is trying to understand Mary. What does John do? Probably, what John does is make a guess (subconsciously). If I'm talking to Mary, I assume that Mary is me, modulo some set of modifications – Mary is like me, and I'll have to make some modifications if it doesn't work out. So, first I assume Mary is like me. Then when I listen to her, it turns out not exactly, so I have to introduce some modifications to get her to be like me. Most of this just happens automatically. It's probably peripheral to the nervous system. You tune into other people in pronunciation and probably in anything else, and you try to find a way of interpreting what they say as what you would've said, and if you can make that interpretation, then you understand what they said, namely, what you would've said.

Sometimes that process is easy, sometimes it's hard, and sometimes it's hopeless, and it just varies all over the place, but there's nothing non-internalist about it. It's always internalist. And you can use any method you like. Most of it is probably subconscious and reflexive, but if you want to try methods like looking things up in a book, writing it down, etc., you can do that too (there are no rules about it). Insofar as I succeed in the task of introducing these modifications, I understand what Mary says, and to the extent that I don't succeed or make the wrong move, I misunderstand what she says (which is often true among people). The point is that positing shared entities has no more explanatory value than positing shared shapes. We don't need any such weird notion to explain communication insofar as we can explain it at all.

Some of these points are made in the literature, but sometimes in a queer way. There's an important paper by Donald Davidson in which he constructs what he calls an 'interpreter'. He says that the basic topic to be looked at is the interpreter. The interpreter is a model of a person in a communication situation, say, John, and the interpreter is getting expressions coming in and is trying to figure out what Mary said. So, the interpreter will give some expression (maybe in a language of thought or whatever you think this happens in) that accounts for what Mary said. And Davidson points out quite accurately that you can use any method you want. In fact, he says, again quite correctly, that interpreting is like finding your way around in the world, therefore anything goes for interpretation. That part is quite correct.

He then goes on to draw very odd conclusions. He concludes that there is no use for the concept of language at all serving as a portable interpreting machine set to grind out the meaning of an arbitrary utterance. So, there's no purpose for any notion of generative procedure, I-language, or whatever. We're led to abandon not only the ordinary notion of language, but we have erased the boundary between knowing a language and knowing our way around the world generally (and he means this conclusion to apply both to linguists who've talked about languages as a portable interpreting machine and to philosophers). Everybody who is talking about languages is deeply wrong because there are no such things. There are no languages. There's no computational procedure for associating form and meaning. It's all a mistake, and the reason is that when we meet somebody, we construct what he calls a 'passing theory'. John constructs a passing theory about Mary, a tentative theory based on what he's picked up about Mary using any device that he can, and that's all there is. There's a passing theory, and once you say goodbye to Mary, you throw it out and wait till the next communication situation.

We have erased the boundary between knowing a language and knowing our way around the world generally, so there's no more any notion of knowledge of language or any notion of language. Since there are no rules for arriving at passing theories (anything goes), we must give up the idea of a clearly defined shared structure which language users acquire and then apply to cases.

There's no such thing as a language. Period. That's in fact the opening sentence of a recent book on Davidson's work, which Davidson endorses as the best exposition of what he had in mind. It opens by saying that there's no such thing as a language, and shows how Davidson got to this point over the years – first the book says there's no such thing as a sentence, and on and on, and it finally gets to the ultimate point that there's no such thing as a language, and the reason is that there are no rules for finding passing theories. Actually, notice that there's something right in those conclusions – we have to give up the idea of a clearly defined shared structure, and that I agree with. There never was any such notion. He puts that in the same category with everything else because of the near universal assumption that it's been shown that you can't account for communication unless you have a shared structure, therefore, it's a big point for him to throw that out along with the idea of a language altogether. And that we should indeed throw out, and that we have never invoked. But not the other part – that we have to give up the concept of a generative procedure that characterizes the form and meaning of utterances, that we have to give up the notion of a language as something that's a state of the brain, that there is no such thing as knowing a language and that these are just parts of finding your way around the world.

Well, the initial statement is correct, that you can use anything you do to form a passing theory. But that's like saying that if you take a motion picture of the world as everything is happening, does it follow that you should give up any effort to do (say) chemistry because that's not everything that's happening in the world? Well, obviously not. To understand what people do when they try to adapt to other human beings (i.e., making a passing theory) is so far beyond the possibility of inquiry that no serious person even thinks about it. That's the kind of thing you write novels about. You don't carry out scientific inquiry about it. It's just too huge a topic. In the natural sciences, there's no such topic as studying everything. You can't do that. There's no such question in the sciences as 'why do things happen?' or 'why are things the way they are?'. Those are questions, but they're not questions in the sciences, because they just involve everything. And if you want to study any topic, what you try to do is find some part of the general mess of things which has the remarkable feature that if you look at that, you will get to some of the basic principles, and ultimately maybe it will slowly build up to getting some sense of the junk that's going on in the world. But that's obvious, and to refuse that in this case is another form of really quite pernicious non-naturalism. I'm sure Davidson will talk about that when he comes here next time, because it's a leading position in his current view of the world. He thinks that's in fact his major philosophical contribution, but it just seems to me to follow from a dramatic form of non-naturalism.

This is a class of arguments to try to show that either language doesn't exist at all, or that it's not in your head if it does exist, and there are other arguments of that type, and they take us into the last topic I want to get into – the topic of meaning and reference. That's the core of semantics. In the last 20 years, there's been actually quite intriguing work done on what are called 'externalist' theories of meaning and reference. This starts with papers by Putnam and Kripke, two quite important contemporary philosophers, at about the late 1960's and early 1970's. Then there's further work by people like Tyler Burge, and by now by thousands of people.

Externalist theories of reference and meaning aim to show that meaning isn't in the head, as Putnam put it, contrary to traditional beliefs that meanings were like concepts or something or the other in the head. These efforts try to show that meanings are not in the head, that they are in the facts about the world and in the social interactions among people. That's where meanings reside. Now, since meanings are, from Frege on, what determines the fixation of reference – the meaning of the word is the method for fixing its reference – then it's going to follow that reference is not in the head, not only in the sense that the thing you're referring to isn't in the head (which, of course, everybody agrees to), but the mode of fixing

reference that I use when I call this thing a table isn't in my head either, it's in the world. It's either in physical reality (which I may not know at all), or in the social world (of which I have only a partial and erroneous grasp).

See, this fits very well with the idea about 'livid' and 'disinterested', that I may not know the meanings of my words. In fact, it will turn out that you commonly don't know the meanings of your words. You'll only know the meanings of the words you use if you know the ultimate answers that physics is someday going to give about reality. And if you know everything about the relevant norms and conventions of your community (which is very unlikely), only then you will know the meanings of the words you use, short of that, you'll only have a partial and erroneous grasp. That's the same conclusion, but it comes from a quite different course.

These externalist theories (which are almost universal and are not even debated anymore) define a notion called 'broad content', which is the externalist notion of meaning – the notion of meaning that's determined by the truth about the world and by the social world in which you live. So, the meaning of the word 'water' is determined by reality. If scientists find out that there's a variety of water called deuterium oxide (D2O), then that's what you were talking about all along, although you didn't know it, because that's part of reality. And if they find some other thing tomorrow, then that's what you were talking about all along. Furthermore, if experts in what's called your community have information about the world that you don't have, then the real meaning of the words you use is determined by their practice, not yours. So, it's going to turn out that, short of accidents, you rarely know the meaning of words and you rarely know what you're referring to. That's going to follow. I'll go through some examples in a moment, but that's the general picture. That's the externalist theory of meaning, which is overwhelmingly adopted at the moment.

The theory is divided into two types – social factors in meaning and real-world factors in meaning. That's the two types of things that determine meaning and reference – all externalist, all outside the mind. There's a big debate in the literature about whether in addition to broad content is there also narrow content. Narrow content is the traditional notion of meaning in which the meaning of the word is in your head. That's narrow content. There's no longer any debate about broad content – that's accepted, that's demonstrated, that's there. But there is a debate whether in addition to broad content there's also a notion of meaning which is just inside your head – kind of an I-language version of meaning – as is the traditional view, in fact, that meaning is in your head. So, when Putnam came out with his article about 25 years ago saying "meaning ain't in your head", that was considered really shocking,

because it had always been assumed that it was, and the argument seemed to show it isn't. So, the question is whether there is also a notion of narrow content, but there isn't any debate about broad content. Jerry Fodor gives arguments saying there's narrow content too, but virtually nobody argues about the non-existence of broad content. The externalist views have swept the field in the past 20 years. We're now into the technical literature on the theory of reference and theory of meaning. There are arguments, but what one has to see is whether the arguments have a natural internalist explanation. Notice that this is going to be wildly non-natural because there are going to be things like 'real meanings' that are just unidentifiable in people's knowledge or behaviour and so on – they're going to be Platonist or whatever. Again, it's striking that there are no comparable arguments given about pronunciation. If you run through the arguments in the literature, you'll find that you can give comparable arguments about pronunciation pretty much all the way through.

Let's start with some of the alleged social factors in meaning. The standard paradigm of discussion that goes back to Tyler Burge has to do with somebody whom he calls 'Bert', and Bert goes to the doctor and complains that he has arthritis. Arthritis is the disease of the joints. Bert, not being educated, tells the doctor, 'I've got arthritis' and the doctor asks 'Where?' and Bert says, 'In my ankle and my thigh.' The doctor looks at Bert's ankle and finds out that he just sprained his ankle and doesn't have any arthritis there, so Bert is just wrong about that. But with regard to his thigh, he isn't just wrong, he misused the word – arthritis isn't the kind of thing you can have in your thigh. Therefore, there is a real meaning of the word, and the meaning is in the social community, and Bert just didn't know it. So, what will happen in this case? Bert will either change his meaning because he defers the doctor's expertise, or he will say, 'I'll keep doing it, because it's my choice.' Does this demonstrate broad content? Does this demonstrate that the meaning of 'arthritis' is outside Bert's head (in this case in the social world)?

Well, there's a perfectly simple internalist paraphrase of all of this – Bert has a certain I-language. In his I-language, 'arthritis' means pain in various places in the body and it includes the thigh and the ankle. He goes to the doctor and the doctor tells him that the word is used differently by professionals, and Bert may decide to pick up that new usage or not, that's all. That's completely an internalist account, and as far as I can see, there's nothing substantive missing from it.

Well, those are the intuitions, and underlying it is the notion of misuse of language, which is extremely common. What do we mean by misuse of language? There is an I-language version of misuse of language – I have some I-language in

my head, and I may well speak in a way that's inconsistent with it, either because the sentence got so complicated that I got lost, or I'm confused by some other factor, or I had too much to drink this afternoon, or whatever. That's misuse of language in an I-language sense. After all, your language faculty is only one component of what enters into what you do. What you do is not a reflection of your language faculty, it's the result of the interaction of the language faculty and a thousand other things. So, it's very possible that what you produce, or the way you interpret something will not be in accord with the actual rules – that's highly possible, and there are many ordinary cases like that. That's one notion of misuse of language, and it carries over to naturalistic inquiry and to the sciences.

Is there any other notion of misuse of language? Not as far as I can see. Every other notion of misuse of language always involves reference to common public language, or some fact about the world or something, and makes reference to broad content. Now, it's true that we talk about the misuse of language all the time, but we also talk about having beliefs about health. It's perfectly okay to keep talking exactly the way we do, as long as we don't get misled about it. Again, that's Wittgenstein's point – there's nothing wrong with ordinary discourse, but don't be misled about it by thinking it has properties it doesn't have – like thinking that if people look alike then there's a common form, or if they talk alike then there's a common language – and all the notions of misuse of language really mean that you're not using language the way some authority structure says you should, or something like that. In the case where we ask if Bert was wrong about his ankle, we have a simple answer. In the case of whether Bert wrong about his thigh – if by that you mean that Bert was not using the language the way the doctor did and the doctor is the authority figure here, then that's correct. If that's what you mean by misuse of language, fine, but that's completely an internalist notion – the doctor has a state of the brain, Bert has a state of the brain, and in this social situation the doctor is privileged, so, if you want the doctor to fix you, you better let him do it his way. There's nothing externalist in that.

That's a typical example of a huge class of cases of alleged social factors in meaning. It's also argued that this goes beyond language, that it goes to concept formation. Burge, for example, argues that to say that Bert has or has not mastered the concept arthritis requires that we refer to a common language in which the concept is expressed with its actual meaning, and so, if Bert hasn't yet mastered the concept of arthritis, it's because there's a real concept, namely, disease of the joints, and Bert has a different concept which he called 'arthritis'.

Again, notice that there's an analogue of this in the case of pronunciation. Suppose Bert pronounced it as 'arthreetis', then is he misusing the language? Is there a real pronunciation, which he doesn't have, in the common language? You could say so if you want, but it would be extremely confusing. What you ought to say is that Bert has a different phonological representation associated with a certain set of concepts. He has a certain pairing of phonological representations and semantic values, and it's not the same as that of the doctor's. How do we know that that's the concept close to the doctor's concept? – because when we try to understand their interaction, we see that they map each other that way. When they're trying to communicate, each of them figures 'I got to reinterpret this other guy' – that's the interpretation procedure – 'and I have to interpret him as saying what I'm saying and I'm going to do it in the minimal way', and these are the points in their respective I-languages that are closest together in terms of phonetic and semantic features. Notice that that procedure may fail – you may pick out something that wasn't really close at all. In that case you misunderstood the other person, which is something that happens all the time. Nothing strange about that. But again, that's a completely internalist account. Suppose I (not the doctor) meet Bert and he complains about his arthritis to me. My first guess would be that he is identical to me, and I think it's a disease of the joints so probably even he thinks so. If I get some evidence that he doesn't, like if he starts complaining about his thigh, then I'll make some modifications in my interpretation procedure for him, and there's no limit to the range in which this takes place and there are no rules for it, exactly as Davidson says. That's a matter of constructing a passing theory, and you can do it anyway you can.

That seems to me an accurate description of communication. There doesn't seem to be anything missing. It doesn't seem interesting because it just says you do anything you can to try to reduce misunderstanding, but that just seems to be a fact. There doesn't seem to be an interesting theory of communication other than trying to work things out in such a way that the other person would be saying what you would've said. That seems to be what communication is about. If it's not interesting, too bad.

Here is another case. Suppose that we say that Peter is trying to improve his Italian. He's misusing Italian now, but he's trying to improve it. What does that mean? What it means on internalist grounds is that he's trying to get his I-language Italian – he's trying to modify it so that it gets closer and closer to a certain range of people (of no determinate boundaries), and he's introducing modifications insofar as he can to get closer to those people in his I-language. That's all it means to say that somebody is trying to improve his Italian. We don't seem to add anything to this

account if we say that there's an entity called Italian which is the common language that he's misusing.

I'll stop at the last case, which is the famous one – Hilary Putnam's 'division of linguistic labour', which is the first example that was given of social factors in meaning. The idea is this. Putnam says that there's a kind of a tree called an 'Elm tree' and another kind of tree called the 'Beech tree', and having grown up in the city like most of us, all he knows is that they're both just big trees with leaves. But presumably, he's using these words with some meaning. He even knows that they're different in meaning. That's all he knows in his head – big tree. If you look up a biblical dictionary, where people are trying to make up meanings for words in the Bible, you'll often find things like 'kind of tree' or 'kind of animal'. They couldn't figure out any more than that. That's all that Putnam knows about Elm and Beech trees too. But there's a meaning to the words. Well, what's the meaning to the words? – the meanings must be given by experts. So, the real meaning of Elm and Beech trees is what the experts mean by it (that's called the division of linguistic labour), and maybe he'll defer to the experts. Like, if you really want to buy a Beech tree to plant in your backyard because somebody said it would look nice in fifty years, you'll defer to the expert and ask him which is the Beech tree and which is the Elm tree, and that's the true meaning, and you don't have it. It's not in your head.

Let's say I don't know the difference between an Elm tree and a Beech tree, and let's say there's an Italian gardener who works in my neighbourhood who only speaks Italian, and we share something – the Latin names for these things. So, I give him the Latin names and he says, 'Oh, that's this', and now I know what a Beech tree is. Does this mean that the Italian speaker with whom I don't share a word is part of my linguistic community? Is Italian part of English? That's a perfectly good case of division of linguistic labour.

Suppose I want to know the meaning of 'angular momentum' and the only guy around is a German-speaking physicist. So, I write it out and he knows that much English, and he writes something back that I look up in the dictionary. Okay, now I've got the meaning of this technical term, and I've got it from an expert who doesn't speak English at all (maybe only minimally), therefore, it must be that we're part of a shared community. In fact, if you pursue this line of reasoning, then everybody's part of the community, because there's only one shared language since you could do this any possible way – and you don't need any of it since it's only a question of how people with particular I-languages choose to accept certain deference patterns, which is a question of authority and its interrelation with

language. Like, I might decide to accept what the Italian gardener tells me or I might not.

In my I-language (like Putnam's), Elm and Beech have different entries but they both have the same semantic features. There may be a general principle about the lexicon that says – if two words aren't *identified* as having the same semantic features, they have different meanings. And these two aren't identified as having the same semantic features, they just happen to not differ in semantic features. So that's presumably the basis for our knowledge that whatever they are, they're different. But that's just a general property of lexical structure, beyond that there's nothing to say. I can go to an expert and modify this peripheral part of I-language – the lexical names for what are called natural kinds (though they really aren't) – that's the kind that you can modify through life at will, and sure we can do that. It's very peripheral to language, and there's nothing more to be said about it. We can answer everything from an internalist point of view in an uninteresting way, but the questions are uninteresting. The questions are uninteresting because there's every possible form of deference relation, authority relation, social interaction, and so on and there isn't going to be anything to say about these topics in general, except 'Do the best you can, and make changes if you feel it helps your life out' or something like that. There's nothing more general to say.

Final comment. For perfectly familiar reasons, there's nothing in this that suggests any problem with ordinary usage – just that we not make mistakes about it. Actually, remember that there was a reason Putnam had. He argues that this externalist view is necessary to make sense of the intelligibility of science. We want to explain that Niels Bohr wasn't talking nonsense when he was using a pre-quantum-theoretic account of the electron, but that doesn't seem necessary either, for reasons we've discussed before.

Lecture #9

I WAS TALKING yesterday about externalist theories of meaning. We're now at the heart of the topic of the theory of meaning and reference. The theory of meaning has two aspects. There's reference – the alleged relation between words and things. And there's meaning – which is supposed to be whatever it is that fixes reference (a framework that comes from Frege). The externalist theory of meaning holds that meaning is not in the mind, so the fixation of reference is not in the mind. And what a word refers to depends on, first, social factors and, second, real world factors. Since the reference of a word depends on those factors, therefore, obviously meaning can't be in the mind. This is by now the prevailing theory. It yields what's called a notion of broad content, i.e., reference. And a mode of fixing broad content, i.e., meaning.

There's then a debate that goes on where it's taken for granted that this has been established. So, if you look at the literature, you'll find in the background the assumption that these things exist, and then the debate proceeds as to whether there's also something called 'narrow content', which would be the internalist version (and more or less a traditional version) – conception of meaning as something that's in the head. So, according to the internalist version – you know what you mean. According to an externalist version – you don't know what you mean because it's determined by the outside world, the future discoveries of science, by societies in which you play a part, etc. This goes along with the idea that each person has only a partial and generally erroneous knowledge of their own language. And language is what it is even if nobody knows it. Similarly, the meaning of the word is what it is even if nobody knows it, and, typically, nobody will know it because the actual meaning will depend on future discoveries of science.

We might first of all ask what this theory is about? A theory is a theory about something. The idea is that these external factors provide us with the real notion of meaning and reference, and then the question of what we mean by *real* arises. Well, there are two possible interpretations of this. One is that we're trying to capture the common-sense notion, what's called Grandma's view – the notion of meaning and reference in ordinary usage. So, that's one possibility – it's the real notion of meaning because that's Grandma's notion. The second possible interpretation is that it's the *true* notion, i.e., the notion of meaning that's accurate about the world, it tells

you the truth about the world. It's the notion of meaning that will arise in the ultimate best science about the world. So, this notion is part of science.

It's never been made very clear which of those two notions is the intended target of the approach, but probably the first one – Grandma's view, or folk psychology – that's what the philosopher is trying to capture in this theory of meaning. Well, it doesn't seem that that could be correct, for one simple reason – in ordinary language, there is no notion of reference as a relation between words and things, therefore there can't be any notion of broad content or narrow content or anything else. There is a notion called *refer*, but that is used for people, like most of the intentional language is – *understand, think, know*, etc. *Referring* is an action that people do in any language I've ever heard of and certainly in English. People refer to things using words or gestures or whatever, but the gestures and the words don't refer, except maybe very derivatively. In fact, you can refer to things without even mentioning. You could say that John was referring to London when he was complaining about the traffic, or when he was telling us what a wonderful performance of Richard II he saw. He didn't mention London or have any phrase indicating London and he could perfectly well have been referring to London. Sometimes you can speak of words referring, but more or less indirectly – you can say "John was using the word 'London' to refer to Paris", maybe out of perversity or something (that's certainly possible), and then we could go on to say "in John's last statement, the word 'London' referred to Paris". Okay, now we have a relation between words and things, but it's highly derivative. The core usage of *refer* has to do with people and what they do, just as all the intentional vocabulary. I've never heard of a language that has any other character.

In fact, when Frege established the subject, he had to make up new technical terms to capture what he wanted, the technical terms 'Bedeutung' (reference) and 'Sinn' (sense or meaning), which are variously translated in various languages. Of course, they are German words, but he explicitly gave them special technical meanings because they don't have those meanings in German, and when they're translated into English, they are technical terms – they have whatever meaning you give to it, there's no debate about what it means. If somebody defines 'momentum', we no longer have a debate about what it means for them because they've told us. We also have no intuitions about technical terms. You can't have any intuitions about (say) 'angular momentum', or 'undecidability' in the technical sense. We can't talk about what it means to you or something. That doesn't make any sense. It means exactly what it was stipulated to mean. So, notions like *reference* and *fixation of reference* (if *reference* is held to be a relationship between words and things),

whatever value the notion has, it cannot be part of Grandma's view because she doesn't have any such notion, although it is an intelligible notion.

Then comes the question – does there exist a relationship between words and things as is claimed? Just defining the notion of reference doesn't establish the relationship (you can define all kinds of notions), you still have to show that the relationship exists, and maybe it does, maybe it doesn't. In any event, it cannot be that the intent or effort of this inquiry is to establish folk psychology. In fact, if it were an attempt to establish the ideas of folk psychology, one would wonder what it has to do with philosophy, because that's a topic in cultural anthropology. Just as the study of people's ordinary conceptions of the natural world is a topic in cultural anthropology and ethnoscience, which is a fine subject and even illuminating.

Let's turn to the second possibility – this theory is an attempt to find the truth about the world. That sounds incomprehensible. I don't know what that even means. What truth about the world are we trying to discover? Maybe one can make some sense of that, but at least an explanation of that has to be given. What aspect of science are we now investigating? If you're doing biology, I could sort of understand what you're doing.

Let's grant that we have some understanding of this, although I don't think it's obvious that we do, but let's grant it for the sake of argument. Let's compare the notions of broad content and narrow content. Let's say we have some phenomenon (say, Bert and the arthritis), and we can give an explanation for it in terms of broad content and also in terms of narrow content. Then the explanation in terms of narrow content wins. It's the one that would be preferred – for purely methodological reasons. One reason is that the explanation in terms of narrow content is clearly part of science – it's naturalistic and it therefore carries no burden of proof. Narrow content is, remember, a scientific question, a question of biology ultimately – to discover what class of linguistic expressions get formed in the child's mind via the generative procedure – and narrow content would involve whatever the semantic properties of this class are. The class of linguistic expressions is a collection of properties – phonetic, structural, and semantic properties – and if indeed there are semantic properties (which we can think of as instructions for use; analogously, phonetic properties are instructions to the articulatory-perceptual system), which is an internalist notion of meaning, and if that notion of meaning can be used to explain phenomena, then we don't even look at other notions (like externalist ones) because this is the methodologically preferable one – because it's part of the natural world, clearly, and it's a notion that arises from naturalistic inquiry, and hence carries no further burden.

In contrast, the externalist approach would carry a heavy burden of justification because it is not a naturalistic concept. It's a concept, in part, of sociology, and it also has the problem that since real world factors enter, the notion of external content depends on all of science – to discover the meaning of a particular word like 'water', you have to in principle depend on all science that is relevant to that. It's a highly holistic notion – so holistic that it's completely comprehensive and includes everything. A concept that in order to be fixed requires knowledge of everything is not a naturalistic concept in any serious sense.

Insofar as social factors are involved, we're pretty much at sea. There's no indication of how social factors are delimited, what are the right communities, and so on. If this is a naturalistic concept at all, it's a very tall order – maybe a concept of sociology or something – and a concept of biology always wins if it is equally explanatory. *[the "Firstly" part is lost here because of some seconds of blank in the audio]* Secondly, the externalist notion involves any kind of social interaction that might be involved with persons deciding to use the word one way or another, and any inquiry into the world that might tell us the truth about what the word actually refers to. Thirdly, the internalist notion appears to be a prerequisite for the externalist one, and that's worth thinking about for a minute. The question is whether you can even present the externalist notion – the notion of broad content – without presupposing the internalist notion. If the answer to that is no, then surely the internalist notion alone suffices to provide a description, and you won't go on to the externalist one, for that reason alone. Well, is that the case?

Let's take Bert and the arthritis again, and remember the way the situation was set up. Bert goes to the doctor and says he has arthritis in his ankle and his thigh, and the doctor explains to him that 'arthritis' is just used for joints, and he doesn't have it in his ankle because he only sprained his ankle. But Bert went to the doctor with some kind of concept in mind. He had a word 'arthritis' that had some semantic features (according to Tyler Burge the wrong ones, whatever that means). He didn't go to the doctor and say 'I have glub in my thigh.' He said he had 'arthritis' in his thigh, and 'arthritis' had a specific meaning, and the doctor had to correct him because Bert *had* semantic features, if he didn't have any semantic features the doctor couldn't have corrected him.

Similarly, in the case of Elm and Beech, they weren't just nonsense syllables for Putnam – in that case he wouldn't have known how to go to an expert. Elm and Beech meant 'big tree' or something for him, so they had some features. In fact, if you look through the literature, you'll discover that in all the cases, the discussion (which is a descriptive account of how meaning is fixed) starts by the assumption

that there's some initial array of semantic features that the terms have (maybe the wrong ones according to this theory). Well, is that eliminable?

[A question is asked by an audience member =] Is it possible to have a notion of I-language in which expressions lack semantic features altogether and you turn to the world or the society to fill in the semantic features?

It's not so simple to construct such an idea. If you have no semantic features at all, how do you approach the problem of adding semantic features? Is it that you've got a concept of semantic features but you just never fill them in? If you do, then there has to be some structure to that concept – there have to be certain things that are possible semantic features and others that are not. This is a theoretically possible account, but it just seems to me hopelessly unpersuasive. You have to assume that in UG, in the child's mind, there is some characterization of the class of possible semantic features, just as there must be a characterization of the class of phonetic features. If there's no initial categorization of that type, you'll never pick up any of them no matter what evidence you use. That's just ordinary logic. Unless there's enough structure to tell you what counts as a semantic feature, you'll never find any of them.

So, we have to assume that the child has some theory of semantic features in the mind in some form or other, but doesn't use it. Rather, the child develops a concept of language in which he never uses those features, and then turns to the actual world and to the future discoveries of science to fill in the features. I don't even know what this means. Obviously, the child doesn't turn to the future discoveries of science to fill in the features. When you say that the child turns to others to fill in the features, how is that any different from saying the child uses the data to add features to the items of the lexicon? That just seems to me a way of restating the internalist picture.

Let's try the same thing on the phonetic side. Suppose somebody came along and said 'I have a new theory of acquisition of the sound structure of language.' The child, of course, has a theory of universal phonetics (has some kind of characterization of the possible sound features), but the child doesn't use them. The child just creates the lexical item without any phonetic features and turns to see what other people are doing to add phonetic features. That doesn't seem to be any different than saying the child uses the data to build up lexical items that have phonetic and semantic features. It's a funny way of saying that simple thing.

There seems no way, as far as I can see, of avoiding an internalist notion in constructing the concept of external content, any more than there would be a way of developing some corresponding phonological concept. You might try to think about it and see if you can imagine a way in which the notion of broad content can even be intelligibly described without presupposing that an operation of this kind is taking place, that is, that the child has looked at data and constructed items which include various properties including semantic properties. Now, it would be intelligible to say the child has the wrong semantic properties, just as it would be intelligible to say the child has the wrong phonetic properties. When I say "intelligible" I don't mean it's the "right" view, but such a view is always based on a prior internalist account based on narrow content, and according to this view the wrong narrow content.

Let's take an analogous argument. Notice that if the internalist account is a prerequisite for the externalist one, if it is simpler (which is obvious), if it's naturalistic whereas the other is not (which is also obvious), then going back to my original statement – if you can give an internalist account, then it wins for those reasons. Now, in the case of Bert and the arthritis, there's a perfectly fine internalist account. Bert went to the doctor with his own lexical item 'arthritis' (maybe he pronounced it differently, like, 'arthreetis'), and the doctor then corrected him. Bert then introduced some modifications in his I-language, and since this is a very peripheral part of the I-language (not like the system of anaphora, which you can't tamper with), Bert could simply decide in effect to drop this linguistic expression and introduce a new word which he pronounces as 'arthritis', and this new word will have a different phonology and different semantics – it will now refer to a disease of the joints. That's just like learning a new word. [*a few seconds of audio is lost here*] Then we can ask why did Bert do it, and then you go into authority structures or whatever. That's a completely internalist account of that interaction, and it's accurate.

The same is true in the case of Elm and Beech. If I go to my Italian gardener and give him the Latin names, and he tells me which is the Elm and which is the Beech, I can decide to amplify my lexicon by introducing new items which include the features the gardener told me about. I don't have to get into the absurdity of thinking that he's part of my linguistic community, which certainly he isn't because he's a monolingual Italian. But it's completely an internalist account. It appeals to nothing more than ordinary acquisition of vocabulary items through experience, in this case through experience plus decision. That doesn't change anything. So, it's perfectly fine. If you run through the literature, you'll find that in every case of alleged social factors in meaning, there is a perfectly straightforward internalist account, that makes use of the social factors, of course, but doesn't make use of any

notion of broad content. Since the notion of broad content requires narrow content as a prerequisite and is far more complex and non-naturalistic, there's no reason to introduce it, at least as social factors. So, at least that aspect of the externalist seems to be eliminable. That's the aspect that had to do with common public language. And that's just as well because it was an incoherent notion anyhow. Anything that appeals to the notion of a common public language, we're happy to get rid of.

There are analogous arguments in other areas. Take Davidson's story about constructing a passing theory. So, we have a communication situation. We have an interpreter, Peter. He listens to Mary's sentences and constructs a passing theory using any information at all to try to figure out what Mary is saying. A passing theory could be anything, maybe it uses knowledge about Mary's background or anything that's relevant. And then this passing theory looks at that evidence and comes out with some interpretation. That's the story.

The question that might be asked here is, does this approach presuppose the existence of, say, a parser? A parser is a device that takes in data and gives some structural account of it – the structural account the internal parser gives is, of course, based on whatever the I-language is internal to the parser. For example, if I hear you speaking, one of the things that I do is assign a phonetic representation, and a division into words and into phrases, and assign meanings and so on, and I do it in terms of my I-language. Of course, that's not the only thing I do. Since first I have to overcome the problem that your I-language is different from mine, so I also have to work out this class of modifications which will allow me to interpret your expressions as some parallel expressions in my mind. But the question is whether the Davidsonian theory (that you construct a passing theory), can even be presented coherently without assuming that somewhere along the line a parser (something that assigns structure to utterances on the basis of the person's I-language) enters into the picture. I won't go into that but you can try it. Try to construct an account of the interpreter that doesn't presuppose that at some stage in the process you are actually assigning structure to the utterances in terms of your own I-language. I will leave it as an exercise to the reader.

An analogous question also arises constantly in the standard picture of fixation of I-language. So, take this picture. This picture, if presented accurately, is presented in an idealized form, you assume that the data is homogeneous – that it's not half Italian and half English, or it doesn't involve contradictory information, like, say, contradictory phonetics. Typically, one says that we're studying the acquisition of language in an idealized homogeneous speech community, where everybody speaks exactly the same way, they pronounce things the same way, have the same meanings,

the same structures, and so on. And we then ask, what I-language would be produced in the child's mind in such a community? That's what this is really a description of.

A very common critique is that this is an illegitimate idealization, because the real world doesn't have homogenous communities (which is true). But the critique that something is an illegitimate idealization, of course, has to be taken seriously. There are illegitimate idealizations. For example, it can be argued that classical economics involves an illegitimate idealization, that it studies market systems without looking at questions of power – that's a critique, and you have to ask whether it's right or wrong. The question is whether in this case it is an illegitimate idealization, and here we turn again to the matter of presupposition and prerequisites – is there a possible account of acquisition of language in real world heterogeneous situations that doesn't presuppose this account? Just think of what such an account would have to be.

Suppose somebody thinks this is an illegitimate idealization. Then my question to him would be this. One, is it possible for a child to learn a language in a homogeneous situation? In other words, is it necessary for language learning that you hear conflicting incoherent data? Is it a necessary feature of language learning that the data that you hear contradict other data? Well, that's absurd, because obviously it's not a prerequisite for learning a language that you hear (say) both Italian and English intermingled. I don't think anybody would claim that. If so, it's certainly possible for a child to learn a language if the data is homogeneous (consistent, in other words). If that is possible, then there has to be some property that the child has in his mind – call it property P – which would enable the child to learn language in a homogeneous situation if it existed. So, we've now established that there is such a property in the mind.

The next question is – does language learning in the normal situation make use of P plus other things? Suppose the child is faced in a real-world situation with conflicting data. Does the child make use of the property P (which must be there), or does it just not use it at all? Well, there are two possibilities – he either makes use of P plus other techniques to get around the fact that the data is conflicting, in which case the idealized model is a prerequisite for the complex situation, or else we would have to argue that the child doesn't use P – in other words, human beings have this funny property in their heads that would enable them to learn language in a homogenous situation but they never use it, that they have all these complex properties like UG and so on, but they never put it to use.

Well, that's not a physical impossibility – an organism might have some complex capacity that it never uses – but it's pretty outlandish. And if you think that this is an illegitimate idealization, you're committed to one of those two views – either it's impossible to learn language unless the data is conflicting, or that a child has the capacity to learn language with consistent data but it's never been used in the history of the human species. Either of those two assumptions is too ridiculous to take seriously, from which it follows that not only is the idealization legitimate, but it is presupposed in all actual work.

Lot of questions in this regard have come up in sociolinguistics. Sociolinguistics studies things like the acquisition of language in non-homogeneous situations, and often sociolinguists claim that they reject the idealizations of linguists. But I think it's clear that they don't reject it, unless they want to commit themselves to one of these two absurd ideas. Unless they want to commit themselves to one of these absurd ideas, they presuppose the linguists' conceptions tacitly, and then that tells you how you should do the work, namely, the reasonable way – you should ask how the property P (the property that enables you to learn language in a homogeneous situation) is used alongside other talents and capacities to deal with the real-world situation.

Similarly, in the case of the interpreter, you'll ask how the parser (which is based on the operation that assigns structures based on its I-language), is used in more complex situations to yield interpretations. The argument here is pretty much the same, except for one difference. The part that's the same says that if you're interested in external content, you'd ask how internal content is used to establish external content – since it's a prerequisite, so how do you use it? The difference is that in these two cases, it's clear that the interpretation does go on – language learning under conflicting data does happen – but it's not at all clear that the established mode of external content goes on. However, if it does – if people actually are picking up vocabulary items and modifying them on the basis of the real world and the social factors – then an account of that will have to presuppose and build upon the notion of narrow content in any event. So, the debate as to whether there is narrow content makes no sense because it's presupposed, unless they want to commit themselves to absurdity. This type of argument applies over and over again, that's why it's worth thinking about it, and I think it's accurate.

In this case, we appear to have a purely internalist account of social interactions that makes no appeal to things like common language. That account is preferable on all grounds, and indeed it seems to be presupposed in the externalist account. So, that would seem to settle that issue. There's a kind of epistemological

priority to the internalist notion – you have to have it before you build up to the externalist one. Just as there's an epistemological priority to the notion of acquisition of language in homogeneous speech communities – you need it in order to build up to the higher notion. And probably (though it's not certain) the same is true about parsers and interpreters. If so, it seems that we're not forced to move to the study of everything in order to deal with the notion of meaning. At least with regard to social factors, it's perfectly possible to keep to a completely internalist account.

I'll come back to the harder question – the real-world factors – in a minute. I think it is harder and more puzzling in certain ways. But let me make clear what I'm not saying – I'm not suggesting that naturalistic accounts are a possible alternative to the standard externalist accounts. Rather I think that the externalist accounts cannot be seriously considered. They can't be seriously entertained, at least with regard to the social factors, because there is a naturalistic account, and, secondly, because the basis for the externalist account – the notion of common public language or community or whatever – has not been clarified, so it's just hand-waving, hence unnecessary. So, it's not that a naturalistic account as an alternative can be placed alongside the standard theories, rather it's the only theory. The standard ones just don't deserve consideration, at least for social factors.

Let me repeat something I said before. I'm also not suggesting that we use the concept 'idiolect' instead of language (or dialect in the sense of a common language). The term 'idiolect' in all this literature is used to refer to the particular style of speech for a particular person in a particular situation, and that notion is also not of much use – it's not an I-language because it was developed in the person's head in the actual world, and the actual world is one of heterogeneous speech communities which involve conflicting data. So, what has gotten into your head is some horrible mess based on property P plus whatever historical accidents that have given this thing in your head. There's something there but it's certainly not an I-language, and the thing that's there is of as much interest to science as the flight of a feather on a windy day is – it's just some mess and it's not worth looking at.

What you want to look at if you want to understand what's going on in the real world is what is called the 'idealized concept' – the concept of I-language as it would arise under homogeneous conditions. And the notion idealized can be misleading – 'idealized' means 'real' to the sciences. Idealization carries you to reality – to the elements out of which the real world is constituted. That's what idealization is for. You can be mistaken – maybe it is an illegitimate idealization – but if you carried out idealization correctly, you're not dealing with an ideal world but with the real world. The reality of the world is those hidden truths that we can

only discover by idealization, i.e., by throwing away complexity that's messing up the situation.

That's in a way the major insight of the Galilean revolution. Galileo, for example, was criticized for studying things like balls rolling down inclined planes. Who cares about balls rolling down inclined planes? That doesn't happen. What happens is balls roll down mountain sides, and so on. And he had to make a propaganda effort to convince people that if you wanted to understand the real world, you had to look at highly idealized situations – like balls rolling down frictionless planes, something which can't even exist. And it's only after that intellectual wrench has been made that you can enter the world of natural science. That wrench hasn't been made in most fields, especially in these fields, which is still prescientific – the understanding that to get to reality, you have to idealize. This is considered quite controversial in the study of humans, and that's another sign of primitiveness.

So much for the social factors, what about the real-world factors? That's a tricky issue. There has been a lot of sophisticated work on it. I'll give a couple of cases. Again, we have to ask the same question – when we look at the real-world factors, are we trying to capture Grandma's view or the true theory? And we have the same problems here too. If we're trying to capture Grandma's view, she doesn't have any views about reference, therefore she can't have any view about the real-world factors. If we're trying to develop the true theory of reference, it's unclear what the topic of study is and it's unclear what the empirical conditions of this inquiry are. It's not like trying to explain the truth about perception or something. Maybe we can clarify this issue by looking at the matter of perception, because, in fact, these arguments have been applied not just to language but to psychology generally, that is, it's commonly believed that all of psychology has to be externalist – has to deal with real world factors. It's widely assumed that all of psychology is crucially wedded to real-world factors.

Take vision. A common view today is that the psychologist studying vision must take into account the real nature of the things being seen. The argument for it goes like this. We have the sentence "John sees a tree" – and that's what's being studied in vision – the fact that John sees a tree, or something simpler like a cube in motion or something. 'See' is what's called a success word – you don't see something unless it's really there. If it's not really there, you may think you see it but you have an illusion. This is not so obvious, but it's claimed that in normal language, 'see' is a success word. And there's something to it. You can't say "I saw a tree" if there wasn't a tree there. You may have thought you saw a tree, but you didn't see a tree. That's what an illusion is, that you think you see something but it

isn't there. Well, since 'see' is a success word, and visual psychology is studying seeing, then it has to be externalist – it has to take into account the thing that is being seen and its nature.

The internalist theory of perception, in contrast, is the study of what goes on between the retina and the brain – it doesn't matter what's out there. Internalist psychology asks what happens when a signal hits the retina and gets interpreted in the brain, irrespective of what caused the signal to hit the retina. But externalist psychology must look at what caused the signal to hit the retina, because you didn't really see the tree unless it was there. For example, if there's a way of stimulating the retina with the same signals that are provided (say) by a cube rotating in space, then you would have in your head exactly the picture you get when you see a cube rotating in space, but you wouldn't have seen it because there was no cube – it was just some mad scientist playing with your retina. So, an externalist theory of perception would insist that psychology must be the study of the relation between things and percepts. Whereas an internalist theory of perception says psychology is the study of the relation between retinal images and percepts.

Notice that an internalist psychology has no problem with the fact that 'see' is a success word – it just says I'm not studying seeing. Seeing is something people do, and I'm studying the way in which they do it. I'm studying the mechanisms that enter into what people do in the real world, and those mechanisms lie between the retina and the brain, and then when I want to talk about how people find their way in the world, I'll add all sorts of other stuff.

Well, it's commonly argued (particularly by Tyler Burge and Martin Davies) that actual visual psychology is externalist. They're interested particularly in the work in the David Marr school. David Marr was a neurophysiologist who pretty much opened up large parts of the contemporary theory of vision by developing models of the visual process that carry it from stimuli up till partial images of percepts, and these models had neurophysiological reality and some successful explanations. So, Burge and Davies and others have argued that in fact if you look at the work in Marr school, it is externalist. Therefore psychology is externalist as it ought to be, and therefore the study of language ought to be externalist too.

But this is just a misinterpretation of the work, and you could see that very clearly if you look at the experiments. The experiments don't involve cubes rotating in space, they involve tachistoscopic images. The standard experiment would be that you have a tachistoscope in which there's a bunch of points of light, and the question is, what different arrays of points of light will lead to a person seeing a cube rotating

in space, and there are all kinds of theories about that. For example, it turns out that if you have four points of light and you give them three successive presentations, that determines a rigid object rotating in space. Your mind does all the rest of the work – it just takes the successive points and forces an interpretation of a rigid object. There's some kind of *rigidity principle* in your head which says that however limited the stimulation (and it can be very limited), it is interpreted as a rigid object moving space. That's in fact one of the big discoveries, but the discovery was not made by having people look at objects through space, it was made by points of light flashing on a tachistoscope, because it was only interested in the relation between the retinal image and the brain.

People have been misled by the talk about functional utility, evolution, etc., in the kind of informal discussion where there are efforts to motivate what you're doing, and, of course, all that talk is externalist. But when you get to the actual experimental work, it's all completely internalist. And, as far as I can see, it has to be. There's no way to do externalist psychology. Externalist psychology would be a study of everything, and we would have to include physics and chemistry and everything else. That's not a topic. So, there's no way to do externalist psychology, and, in fact, nobody does it. They do internalist psychology. They may describe it in ways that mislead people (who maybe mislead themselves), but there's no possibility of doing externalist psychology.

Notice again that if there is an internalist account of perceiving a cube rotating in space (as there is), then we don't look for an externalist account, because the internalist account meets the burden of proof, it's strictly part of science, is obviously far simpler (it doesn't require the study of trees, say), and it's a prerequisite – you can't develop the externalist account without some account of what happens between the retina and the eye. Even if you're interested in studying the relationship between trees and percepts, somewhere along the way you'd have to study what goes on between the retina and the percepts. So, the internalist account is a prerequisite (and obviously a prerequisite in this case), therefore it wins. And since it works anyway, there's no further interest in studying externalist psychology.

Let's take a look at the special branch of psychology that studies language and ask whether there is any more sense to an insistence of real-world factors in reference and meaning. Here the arguments again go back largely to Putnam and Kripke in that same collection of articles that I'd mentioned, which, about twenty years ago, really set the field on a new course. Incidentally, they have a lot of not only well-argued but very useful material in them. They were quite important, although I don't agree with these particular conclusions.

The standard argument is what's called the Twin Earth argument. That's the paradigm around which discussion proceeds. Putnam proposes the following thought experiment. We on earth have a word called 'water'. Let's pick a person on earth, who he calls Oscar, who has a word 'water', and that word refers to a particular thing which scientists have told us is H2O. So, the word 'water' refers to H2O. So, that's a fact about our language, scientists told us that. Okay, that's the real-world factor.

Now imagine another place far away in some other universe in which there's a duplicate of Oscar called Twin Oscar, and since he's exactly the same as Oscar, he's also going to have the word 'water', because his brain and everything about him is the same as Oscar. But suppose in that universe there is no H2O but there is another thing with a chemical composition XYZ which happens to look exactly like water, and that's what Twin Oscar is referring to by 'water'. So, it therefore follows that meaning isn't in the head, because Oscar and Twin Oscar have exactly the same thing in the head (by assumption), and if what fixes reference is something that's in the head, then they would have to have the same reference, but they don't because the words refer to different things. So, we have therefore established that meaning isn't in the head. That's the paradigm that's argued.

Well, is it really convincing? Firstly, notice what the thought experiment is – it's exploring our notion of reference of words. It's saying that under these conditions, we would say that the word 'water' for Oscar refers to H2O, and the word 'water' for Twin Oscar refers to XYZ. That's what you're supposed to be persuaded of. To be persuaded of that, you explore your own intuitions about what words refer to, and you discover that your intuitions have this property. That's the argument. It's not an empirical inquiry or something, it's an introspective inquiry into your own intuitions about the word 'refer' that holds between words and things.

Well, point number one is that we have no such intuitions. We can't have any intuitions because the concept of 'refer' that's holding between words and things is a technical, invented concept. If somebody explains it to you, you can understand it, but it's not a concept you have intuitions about any more than you have intuitions about 'undecidability' in the technical sense. If you're studying metamathematics, and you get to undecidability proofs, and you start using your intuitions about the word 'undecidability', you better take up another field, because you're not supposed to have any intuitions – 'undecidability' means whatever that guy standing up there says it means, and you have no intuitions about it – as a matter of logic.

Similarly, if ‘refer’ as relationship between words and things is not part of ordinary usage of language, then you can’t have any intuitions about it, and as far as I can see, it’s not part of ordinary usage. There is a word ‘refer’, but that has to do with what people do. So, problem number one is that you can’t have any such intuitions. That doesn’t mean that people don’t make judgments. In fact, everybody in the philosophical literature makes the same judgment, namely, that ‘water’ refers to H2O for Oscar and it refers to XYZ for Twin Oscar, so they’re doing something – but it must be illegitimately bringing in intuitions about the use of the word ‘refer’ as something that people do to a situation where it’s inapplicable.

Suppose we use the actual word ‘refer’ – so, in fact, Oscar is referring to H2O and Twin Oscar’s referring to XYZ – then that is correct, but that doesn’t get you anywhere, because if Twin Oscar landed on earth and pointed to a glass of water, he would now be referring to H2O. There’s no paradox – yes, he’s referring to XYZ up there and he’s referring to H2O down here, so that’s no problem. The ordinary use of the word ‘refer’ doesn’t get us into any problems at all. We would say, in fact, that up on twin earth, Oscar was referring to XYZ, because that’s the stuff they have up there, and when he comes down here and asks for a glass of water, he’s referring to H2O – nothing problematic about it, no real-world factors in meaning are involved.

The argument that meaning isn’t in the head only goes through if you’re talking about the relation between words and things – the argument only goes through if you accept the externalist story that when Twin Oscar comes down to earth and points to a glass of water and says ‘water’, he’s wrong, because ‘water’ for him refers to XYZ and this is H2O. That’s the intuition you’re supposed to have. Well, that intuition certainly doesn’t refer to the word ‘refer’, because there’s no doubt that he’s referring to H2O when he comes here. So, the intuition can only relate to the word ‘refer’ holding between words and things, but we can’t have any intuitions about that. So, it seems to be an impasse. It seems that the argument collapses. And whatever conclusions people are drawing must be some illegitimate interaction between a technical term and an actual term about which they do have intuitions, but the wrong ones in this case.

Let’s say that Twin Oscar comes down to earth and sees a glass of water and right next to it he sees Sprite.

In the U.S. there’s something called Sprite. It looks exactly like water but it’s something else. It’s made that way so you’ll pay money for it instead of just drinking a glass of water.

So, suppose there's a glass of Sprite over here and a glass of water over here and Twin Oscar comes down to earth and he points to the water and says 'I'd like a drink of water' and he points to the Sprite and says 'I'd like a drink of water.' In the first case, he is referring correctly, he's not making a mistake. But I think in the second case, he is making a mistake, that is, if he points to Sprite and says 'water', he's making a mistake. But if he points to H2O and says 'I want a glass of water', at least my intuition is that he is not making a mistake, that there's a crucial difference between those two cases. On the externalist account, he's making the same mistake in both cases, because 'water' for him means XYZ, so it doesn't mean H2O and it doesn't mean Sprite. But that's highly counter-intuitive.

Let's proceed further. Suppose Twin Oscar is a kind of Frankenstein's monster who was created this instant to be a molecule-by-molecule duplicate of Oscar. Now, Twin Oscar has had no experience at all because he was just created. So, it is not the case that his word 'water' refers to XYZ, because there has been no XYZ in his environment – he had no environment. Therefore, if Twin Oscar now says 'water', he's not referring to anything. If he comes to earth and points to a glass of water and says 'I'd like a glass of water', it's just incomprehensible. He's not referring to anything. There's no reference. That's the conclusion we're led to.

Since we're playing with thought experiments, suppose there's a law of nature not yet discovered which works like this – every second each person on earth is replaced by an exact duplicate from outer space and is instantly transported here and that person continues life until the next second when he's replaced by somebody else. So, we're constantly being replaced by duplicates, and we would never know it because each duplicate is exactly like us. You can imagine a world where that happens. Now, in that world, nobody would ever be referring – life would go on exactly as it does for us, but nobody would be referring to anything, because they've never had any experience at all, so their words don't have any reference.

When you get into absurdities of this kind, you quickly look back to see what went wrong, and I think what went wrong is the attempt to have intuitions about something we can't have intuitions about, namely, an invented technical term. Therefore, when we try to have intuitions, we go astray, and the intuitions are going to lead us all over the place, and to crazy paradoxes, and so on. I think that's exactly what's happened in the twin earth case.

Now, it's interesting that Putnam picked the word 'water' here. Why did he pick that word and not a host of other words? For example, he could've picked 'earth', 'air', or 'fire'. Those are worth putting in a category for familiar reasons.

They were traditionally viewed to be the constituents of the universe by the ancients, and they regarded all these as on a par (of course, they weren't speaking English, but some counterparts of those words in their languages). So, how come Putnam didn't use the word 'air' or 'fire' for Oscar? Well, the answer is obvious. There's nothing like H2O that characterizes air. In fact, there isn't any such thing as air. There's nothing that characterizes fire.

Notice that we have no problem referring to fire, like, there can be a fire burning in my fireplace, I can refer to the fire, just as easily as I can refer to a glass of water. I can refer to the earth, I can refer to the air, say, the air is polluted today. There's no problem referring to these things, it's just that science doesn't have any concept that's close to them. But is the fact that science has a concept that's close to water of any relevance? It's not obvious why that's relevant. What relevance is it to our speech that there's a notion in the sciences which is not too far from what we call 'water'? In fact, it isn't the same thing. Like, this glass of water is obviously not H2O. In fact, scientists have now discovered that if you have pure water, that's not H2O either – it's a mixture of H2O and D2O (heavy water), and it's a mixture in the proportions of about 6000:1 or something like that. So, according to this story, since the reference of Oscar's term 'water' is what sciences tell us water is, then in fact Oscar wasn't referring to H2O at all – he was referring to a mixture of H2O and D2O in in the proportions of 6000:1. And if in some other universe you have a different mixture (let's say the proportions are 200:53), then Oscar was referring to a different thing.

Notice that this is all totally sub-perceptual – there's absolutely no way of telling the difference between water and heavy water, except by sophisticated experiments. What Oscar is referring to, according to this story, depends on the proportions of D2O in the samples of water in his universe, which, again, is a highly crazy result, and if you push it further it will lead you to further complexity.

To get back to where I was, the question is – why care what scientists tell us about the thing that they call 'water'? Scientists have a thing they call 'heavy water' (D2O), so does that affect my usage of 'water'? – No. It's just a homonym. It just happens that when you do science, you usually use words from natural language because it's too much trouble to make up words all the time. So, you use familiar words from natural language but give it a different meaning. So, people use the word 'momentum' but they give it whatever meaning they want. We use the words 'empty category principle' and we give it whatever meaning we want. The fact that linguists use 'empty category principle' to mean a certain thing doesn't influence the meaning of the word 'empty' in ordinary language, obviously. Similarly, the fact that

physicists have used the word ‘water’ to distinguish, say, ‘heavy water’ (D2O) from what they, in fact, call ‘light water’ (H2O) has no implications whatsoever for what the word ‘water’ means in English or for Oscar.

Putnam selected ‘water’ because he’s really interested in the question of intelligibility in science – how can we understand scientists from 500 years ago unless we assume that they were referring to the same thing that we were referring to? Therefore, it must be that they were referring to what we have discovered to be water, or what some future scientist will discover when he finds that we’re wrong. But that’s just a wrong theory of intelligibility, for reasons I mentioned earlier. There’s no more relevance to the term ‘water’ in this thought experiment than to ‘earth’, ‘air’, and ‘fire’.

Now, the importance of this is that ‘water’ is what’s called a ‘natural kind term’, and the externalist theory of reference that’s being built up by Kripke and Putnam and by everyone else has to do with natural kind terms and with proper names, like, say, ‘Oscar’. The purpose of this theory is to explain the reference and meaning of, basically, names – proper names (people, cities, etc), and natural kind terms, which are the names for the stuff in the universe. But the problem is that language doesn’t have natural kind terms. The terms of language don’t pick out the constituents of nature, except by the most amazing accident. That’s not what language is about. That’s not what it’s for. That’s not how it develops. It couldn’t possibly be the case that our minds are so attuned to (say) quantum theory that the terms that enter into a child’s lexicon are the quantum-theoretic terms for the constituents of the universe. That’s obviously not true. There are no natural kind terms in natural language, at least if natural kinds are the kinds of nature.

Let me just summarize this. The real-world component of reference is an interesting question – it’s an interesting question to try to figure out whether indeed there is a real-world component that enters into reference. But the arguments that are given (this is the key argument and there are many variations) just don’t seem to show anything. They start with an illegitimate import of intuitions from the natural word ‘referring’ to the technical word ‘refer’ (which just means what you say it means), and having made that illegitimate move, you quickly enter into a paradox, as happens constantly when you make illegitimate moves. I’ll give some other cases as we proceed.

Though this is an appealing idea, it just doesn’t seem sustainable. There seems no reason to believe that real-world factors enter into either reference (in the sense of what people refer to) or meaning (in the sense of what fixes what they refer to).

Though an interesting idea, it doesn't seem to apply to natural language or to psychology in general. If that's the case, then the externalist program is without value. It's sort of intellectually interesting, but essentially without value. That's another controversial conclusion and therefore you ought to think about it, because it's virtually dominant in analytic philosophy without any question, though unsustainable in my opinion.

What I want to get to next is a tradition that starts with Frege and comes up till today, that provides, in a way, a more fundamental reason for believing in common language and reference in this sense and so on.

Lecture #10

LOOKING AT EXTERNALIST views, these have, first, a real-world component for meaning, and, secondly, a social interaction component. That's the character of prevailing externalist doctrines. According to the real-world component, a word or an expression picks out the essential nature of the things to which it refers, and this essential nature is often to be discovered, but that's its reference – its reference is to things – by virtue of their essential nature.

The social component says that words get their meaning in a common language through social interaction, and we don't know the right meaning of the word but we learn it in the common language. The internalist view, on the other hand, says that a word or an expression is just a collection of its properties. So, a linguistic expression is just a collection of properties – phonetic, semantic, and others, and one of the ways in which languages differ is in this association between phonetic and semantic properties (Saussurean arbitrariness). When you learn a word, you pick up the association (and if Jerry Fodor is correct then that's about all you learn) and maybe a few options as to how the semantic features would be there).

These are two different pictures, and what I argued this morning is that the phenomena that are addressed in terms of broad content are real enough, but they can be accounted for perfectly straightforwardly, if there is any account at all (which there often isn't), in terms of narrow content – an internalist account – which is, in fact, tacitly presupposed in the externalist approaches anyway. In terms of what's presupposed in the externalist approach, there's already an explanation, and therefore it doesn't seem like there's any need to go on to this more complex and basically non-naturalist doctrine.

Now, it doesn't mean there are no questions left here. There are plenty of them left. With regard to the real-world component issue, we're left with a very interesting question. The question is – is there a relation R between an expression and a thing in the world where R has the stipulated properties of reference?

I was arguing this morning that there's no such concept in natural language, but that doesn't mean that there might not be such a concept in a theory of language. In general, in a theory of language, we're going to have notions that don't exist in

the language itself, and as long as we don't get confused about this, we may turn sensibly to ask a question whether R exists.

With respect to the social interaction component, if you look at the descriptions by Tyler Burge, I think they're quite accurate. So, Burge in his articles gives detailed and often rather perceptive accounts of how a group of reasonable people who are seeking some sort of a common ground might modify their usage. In internalist terms, that would mean to drop certain linguistic expressions and pick up others (and often the changes would be new semantic features or something), and they might decide to do this under many circumstances – with deference to people who know more about some topic and therefore have better decisions about what are useful concepts, and so on. If you look at his articles, I think that's basically what he describes, and I think things like that happen – reasonable people do adjust their concepts and defer to others and try to find common ground and so on – but it's not a matter of finding the real meaning of the word in the common language. There's no such thing as common language. The reason nobody can make a sensible proposal about a common language is that there's nothing there to seek.

And as for the real meaning of words, it doesn't mean anything – one of the aspects of I-language that's readily modifiable is the Saussurean arbitrariness (the association of sound and concept), and you do that all the time whenever you're learning a new word. And in certain situations of social interaction of the kind Burge describes, it might be entirely reasonable for people to explicitly decide to introduce words slightly different from the ones they had (like Bert and arthritis). Many of these accounts are perfectly reasonable and I don't mean to question them, but I just say that they're not saying anything about language – they're telling you something about how reasonable people might behave in certain circumstances with regard to modifications of their I-language. Nothing is added to that description, as far as I can see, by bringing in an externalist theory, except confusion. And it's purely an internalist description of it. It's not a description totally in terms of I-language because it makes reference to people's interests, goals, decisions, etc., but if you're talking about all of human life and no just language, it's very unlikely that there's going to be much to say about this topic, that is, there will be particular things to say about particular cases but virtually nothing of any generality or any theoretical interest, because it's verging towards something like the study of everything – and we know we can't say anything about that. These descriptions are like descriptions of how people live their lives. These are interesting topics, but the kind of topics you write novels about, not make theories about. Crucially, it does not lead to any such exotic notion as common public language.

Well, there's another familiar argument for common public language, which is a classic one, and I want to give that one, and it also relates to these other matters about reference. This is the one that goes back to Frege, and this one is implicit, at least, in the work of non-Fregeans like Kripke and Putnam, who were explicitly breaking from Frege in claiming that meaning isn't in the head.

The Fregean story goes like this. Take *Sinn and Bedeutung*, a classic work. We begin by assuming what he calls a 'common store of thoughts', and Frege, without arguing about it, simply says that one can hardly deny the existence of a common store of thoughts (which is a common human possession). No further discussion. So, one can hardly deny that there is a common store of thoughts which is a common human possession, and therefore we assume it. Then we move to a common public language in which these thoughts are expressed, and although meanings are in the mind, they are objective things – everybody's got the same one. There is something in the individual mind, but that's something like an image or something like that.

And the third point, which could be theoretically dissociated from these, is to present a description of the common public language in terms of the notion of reference, which is a relation that holds between expressions and things generally ('Bedeutung' for him), and a mode of setting reference ('Sinn' for him).

As I mentioned, Frege had to make up technical terms for this because the German words didn't mean that. But that's okay, these are technical terms. In other words, he is answering that question up there positively, that there is such a relation R. Well, that's the picture. And moving in that way, we get a kind of argument for common public language (given that it can't be denied that there is a common store of thoughts).

Well, let's start from the beginning. Maybe the story is right, but it surely is not as obvious as Frege claims, that is, it certainly can be denied that there's a common store of thoughts. In fact, it had been denied quite plausibly long before. It was denied, for example, in the 18th century critique of the theory of ideas – the theory of ideas was sort of the big issue in the 17th and 18th centuries. By the late 18th century, an influential and important critique of the theory of ideas was developing, based on the notion of what became in the 20th century 'the ordinary language critique' – that this was simply giving a wrong grammar to the expression. Notions like "John has a thought" or "John has an idea" are what Gilbert Ryle 150 years later would call a 'systematically misleading expression' – an expression which looks as if the terms in it are referential. It looks as if "John has a watch" or "John has a

diamond", but that's a faulty analogy, because "John has an idea" or "John has a thought" just means "John thinks", and if "John has the idea X" it means "John thinks X".

As far as I know, this critique was first made by the French encyclopedist Du Marsais. Later, it became famous with Thomas Reid, the Scottish common-sense philosopher, who was more in the centre of the tradition. That was a major part of the philosophical tradition and certainly a plausible view. In the 20th century, with Ryle and Wittgenstein and so on, this became a cliché. Maybe the critique is wrong, but it doesn't seem obviously wrong, and if it is right, then you certainly can deny that there's a common store of thoughts, because there aren't any thoughts or ideas at all – people think, they think more or less alike, and that's all there is to it. One can certainly take that view. And to go beyond that and say that 'people think alike, therefore there are thoughts that they share' is itself no argument. It would be like saying that if people look alike, then there must be shapes that they share, which is obvious nonsense. So, a further argument is required to show that there is a common store of thoughts.

Now, of course, there will be domains where there will be a common store of thoughts – because you create them that way, like in mathematics, which is what Frege was interested in. There it makes sense to say there's a common store of thoughts, and if you're a Platonist then they're actually there, and if you're an intuitionist then you construct them or whatever, but there you can say there's a common store of thoughts.

However, in the study of people in the natural world, you can't make that assumption. You have to show that there's a common store of thoughts, and it's hard to see how one could do that. So, the very first step seems highly dubious. If you don't make the first step, you obviously won't go to the second step. If you do make the first step, the second step doesn't follow automatically, but it certainly doesn't follow if you don't accept the first step. And there's no reason to accept the first step unless somebody offers a reason. And the only reason that's offered is basically the grammatical analogy, and that obviously isn't enough, or the fact that people think alike – again, not enough.

Let me just say that there's something a little different between questioning the notion of common language and questioning the notion of ideas and thoughts. Common languages are proposed as entities, and there we have very solid grounds for rejecting the proposal. It's not that nobody can tell us what they are, it's that we know why there aren't any. There's a good theoretical understanding of the issue

which explains exactly why there aren't any such things. Why talk about common languages like talking about absolute regions? So, in that case, rejection of the proposed entity is based on pretty firm grounds.

Now, in the case of thoughts and ideas, nobody knows anything. Basically, nothing is known about the topic of thinking, having ideas, and so on, therefore one is not on strong grounds in rejecting the proposal that there are thoughts and ideas – maybe there are. Since there's just no theoretical understanding of it in any terms, you can't say these are the wrong terms. All you can say is something weaker in this case, namely, that there doesn't seem to be any substantive argument and hence a big gap is to be filled before we could proceed. And given that there's no argument for it, one would obviously take the more limited position that you don't accept it – you reject them until an argument is given, even if you have no reason for rejecting. In the case of common language, there's a reason for rejecting the argument, therefore it's a stronger position. Whatever the answer to this question is, we are on firm grounds in rejecting the notion of a common language, and we're on good grounds, if not solid, for rejecting the notion of thoughts and ideas – the good grounds are that there's no argument for adding to our ontology those entities, and therefore we will reject it even though we have no better theory in some other terms, because there's no theory.

What about the third point? Is there a relation of reference that holds between words and things, and is there a way of fixing that reference? Is there something like Bedeutung and Sinn? That's the question. Now, if you look at the best work on reference, from Frege to the present, there's a characteristic move that's made, which is made too fast. You see it, for example, in one of the best books on the topic, by Gareth Evans, *The varieties of reference*. He begins the exposition by making a statement that's absolutely true, namely, that people use words to refer to things. He says, 'Look, people use words to refer to things and that's what we're studying – reference'. He then makes a move from 'people use words to refer to things' (like, "John used the word 'London' to refer to Paris") to 'words refer to things', i.e., the existence of R. And there's no mark of that transition. It's just one paragraph where we have 'people use words to refer to things' and in the next paragraph we have 'words refer to things', and then we're off and running on a theory of reference and meaning. But that's a transition that has to be justified. You can't just make it. And I think you'll search in vain for any justification for it. In fact, it's usually so tacitly assumed that people even think that they have intuitions about the technical notion of reference – as in the twin earth discussion – which is what Dan Dennett once called 'intuition pumping' (pumping up your intuitions to try to figure out what you think about these things). But even if the technical concept exists, you can't have

intuitions about it, but it is so taken for granted that we have intricate discussions about what our intuitions are about this invented technical term. Well, that won't do.

Now, we have to be careful not to confuse ourselves about another issue. Suppose a linguist is working on some theory of semantics and decides to set up things which he calls 'semantic values', and words have a relation to semantic values, and that relation is, let's say, the relation R. In fact, R will then be assigned the properties of reference. So, 'John' will have a relation R to a certain semantic value. So, we now have some domain D of entities called semantic values, we have a relation that holds between expressions and semantic values, and then you can do a lot of work. You can ask that if you put together expressions in a certain way, how do you get things in D that you're referring to. Or, for example, you can develop a kind of reasonable theory of anaphora, meaning relation between a pronoun and an antecedent. Let's take the examples I gave last time.

(1) He thinks John is a genius.

(2) John thinks he is a genius.

(3) His mother thinks John is a genius.

We know the following facts. In (2), the pronoun 'he' might be referentially dependent on 'John' (whatever 'John' refers to, the pronoun refers to; it's not necessary but it's possible). The person who uses that sentence could be intending a common reference to the pronoun and 'John.' Similarly, in (3), the person who expresses this sentence could be intending a common reference to the pronoun and 'John'. But in (1), no, the person could not be intending a common reference.

It could turn out that 'he' in (1) refers to John, like if maybe there's only one person in the world. Or maybe some other John was referred to in an earlier discourse and 'he' is picking up the reference from that, and so by transitivity it will be the same reference as 'John'. But we know that 'he' is not picking up its intended reference from 'John' in (1).

So, we know different things about these sentences. Those are the facts about language. The theory of anaphora, or binding theory as it's sometimes called, attempts to explain these facts, and a possible way to explain them is by setting up a domain D with the relation R, and giving semantic values to the relevant terms, and to say that in this kind of configuration (which you then proceed to define) the pronoun 'he' can't pick as its reference (a thing to which 'he' is related by R) the

element in D to which 'John' refers. That's a way of building up a theory of anaphora. That's perfectly sensible. Maybe right, maybe wrong, but perfectly sensible, and it's on a par with some theory of phonetics.

Notice that it's all totally internalist. We haven't got anywhere near the world yet, and the world doesn't even have to exist. The domain D is a set of syntactic objects. They're called "semantic" values but you don't want to be confused about that, though I think many people are. The semantic values have nothing to do with semantics – they're syntactic entities. The domain D is a collection of syntactic objects – formal objects – which this theory claims are in the mind. Semantic values are on a par with (say) phonetic features.

Phonetic features aren't in the world, they're in your head. In the world there's just noises, and the array of phonetic features gets converted into noises via the performance systems, and the noises get interpreted in terms of phonetic features, again, by virtue of the performance systems. But nobody would confuse phonetic features with noises. Like, if I write down 'aspirated labial', nobody thinks that I just wrote down a noise. I only wrote a symbol for a collection of syntactic objects – the phonetic features – and universal phonetics will be a theory about what these objects are, and, in general terms, it's a syntactic theory.

We use the word 'syntax' now in the broad sense to include anything about mental representations (there's also a narrow sense which refers to a part of that theory, namely, the part that involves phrases and stuff like that). Then phonetic features are syntactic objects and the domain D is syntactic objects. To go from phonetic features to the outside world is no trivial matter – you need a theory of articulation, perception, etc., and it's hard to work it out. To go from D to the outside world is certainly no less hard, in fact it's not obvious that it's possible at all. The postulation of the set of phonetic values is justified on the usual best theory terms – that is, a person who proposes a theory of distinctive features is, in effect, claiming that the way to understand speaking and hearing (articulation and perception) is to set up this level of syntactic representation and to relate the level to noises by way of universal properties. A phonetic representation's interpretation is language-independent. Once you get down to phonetic representation, everything from there on is outside the particular language in question, and if that's not true, you didn't get down to phonetic representation. It's an interface between the language faculty and other systems. That's the meaning of the notion. If you fail to get to the point where everything is language-independent from that point on, you know you didn't have the right theory of phonetics.

And a theory of D would have to have the same property. It's an interface. It's the point at which the language faculty interrelates with other faculties of the mind, like conceptual faculties, intentional faculties, etc., and the interpretation of D by these other faculties must be language-independent if this is a valid description of the interface. Well, it might be. The proposal should be that D and R are on a par with what's called the Phonetic Form (PF) and some relation P that one might set up, relating expressions to PF. A part of the path from linguistic expressions to PF would involve lots of such relations, but let's pick one of them and call it P. They're all totally internalist, they're all syntactic, they don't in themselves have anything to say about the world, they don't even require a world to exist (all that could be happening is your head exists), and these theories would work exactly the same way.

Well, is this the right way to move? Maybe. Incidentally, what I'm saying now applies to every form of formal semantics that is set up – model theoretic semantics, discourse representation semantics – anything they know about has these properties. It sets up a domain D of syntactic objects which have nothing to do with the world as far as we know, though they may – a plausible argument one could give is that the syntactic representations in D (something about possible worlds) is more transparently related to the actual world than linguistic expressions themselves are, that when you get down to D you have a kind of transparent relation with the world. That would be like saying that when you go from abstract phonological features to phonetic features, you now have a more transparent picture of the world. That amounts to saying that it's a real interface – that there's going to be a language-independent interpretation, which is the property it has to satisfy. That's what it means to say 'transparent relation'. And maybe, but maybe not. Some reasons to think that the answer is no come about very quickly if you start playing with these sentences.

So, suppose instead of 'John' we have 'the young man'.

(4) He thinks the young man is a genius.

(5) The young man thinks he is a genius.

(6) His mother thinks the young man is a genius.

The relations of anaphora remain the same. Therefore, if our theory of anaphora is formulated in terms of R and D, what we'll say is the same thing – 'the young man' picks out an element of D (call it *d*), and we state the anaphora theory exactly as before in terms of the relation R (which we read *reference*), and the pronoun picks out a certain element of D as its reference. So far so good.

Suppose we replace ‘the young man’ by ‘the average man’.

(7) He thinks the average man is a genius.

(8) The average man thinks he is a genius.

(9) His mother thinks the average man is a genius.

Notice that (9) has a certain oddity about it, which has to do with stress – there are some funny facts interacting here about what’s called backwards anaphora with regard to lack of stress. If you don’t stress ‘the average man’ in (8) you also get funny relations, but that we can abstract away from. In fact, we could just look at the first two in case the third sounds funny (which it should because it’s adding an extra property). (7) and (8) have the same anaphora relations, and therefore the same interpretation – ‘the average man’ is related to some element *d’* in D by the relation R, and we therefore give the same theory of anaphora.

And it doesn’t have to be a definite noun phrase, it could be a name too. Suppose we call ‘John Doe’ the name of the average man. If there’s anything to the notion of a common language, this is in it – that’s the standard convention.

(10) He thinks John Doe is a genius.

(11) John Doe thinks he is a genius.

(12) His mother thinks John Doe is a genius.

The same relations of anaphora hold – ‘John Doe’ will get *d’*, and we explain ‘he’ the same way.

Well, what’s the point of this exercise? Remember that the point of moving to D was supposed to be that we’re getting to a transparent relation to the world, i.e., a relation to the world which will have a uniform language-independent interpretation within things, and we don’t seem to be getting there at all, because nobody believes that ‘John Doe’ and ‘the average man’ pick out things in the world. I mean, we understand them alright – ‘the average man’ means add together all the men and divide by their number and so on, but there’s certainly no thing in the world to which ‘the average man’ refers, or to which ‘John Doe’ refers, and nobody is confused about that, I hope.

Notice that the theory of anaphora is at the core of what is called the logic of language – the relation of variables to what binds them and so on – and if you extend it on, you get right into quantifiers, and anything that could be called a logic of language is going to have exactly these properties. And it always works exactly the same way if the terms that you're picking out might have reference and if the terms you're picking out certainly have no reference (where I borrow this term 'reference' pretending we understand it). So, in other words, the internalist justification, such as it is, for setting up the domain D seems to be absolutely irrelevant to the relationship to the world. In that respect, it is quite unlike PF. The move to PF really does take you to something like a transparent image of the actual world, whereas this doesn't seem to take you anywhere. The same happens if we turn to other expressions. So, take, for example:

(13) It brings good health's own rewards.

(14) Good health brings its own rewards.

(15) Its rewards are what make good health worth striving for.

If you think about these three expressions, they have exactly the same formal properties, and we have the same result. In (14), 'its' is anaphorically dependent on 'good health'. In (13), 'it' cannot be anaphorically dependent on 'good health'. In (15), again, 'it' can be anaphorically dependent on 'good health'.

So, we have exactly the same relationship, but we're going to explain the phenomenon in terms of the relation R and the domain D. We will set up an element of D, and since it's a syntactic object we can give it any properties we like. We'll put into D an entity *g* to which the phrase 'good health' bears the relation R, and then we'll get the same theory of anaphora. Again, I doubt that anybody is confused into thinking that one of the things in the world is 'good health'. To understand these sentences, you don't have to assume that among the various things in the world is 'good health' which has these properties, and it would obviously be a mistake to make any such move. Nevertheless, the relations of anaphora and dependence work exactly the same way.

To give a final example, take the sentences:

(16) There was a flaw in the argument.

(17) The argument was flawed.

For all relevant purposes, these are synonymous by assumption. But anaphora works differently for them.

So, try adding "and it was quickly found" to both sentences and ask about the reference of 'it'. In (16), 'it' has anaphoric properties (as a pronoun usually does) – 'it' picks up its reference from 'flaw'. In (17), there's nothing for 'it' to pick its reference from – 'it' must be referring to some other thing (like, the chair was quickly found or something like that). The way we handle (16) in the internalist theory of formal semantics is to say that 'flaw' bears the relation of reference to some entity in D. In (17), for purely structural reasons, there is no word that can have the relation R to some element in D, therefore we can't run through the theory of anaphora for (17). That's fine, except that we now have the case that two expressions which are in fact synonymous (and if you look closer, they may even have the same underlying representation at the relevant level, as has been proposed) have different relations to the domain D, as shown by the same anaphora facts.

Again, that seems to show us that we're not getting anywhere near semantics. Maybe we're getting farther away from it, but we're certainly not getting anywhere near it. Semantics is the study of the relations between languages and the world. It's not the study of syntax. It's not the study of relationships internal to particular systems of mental representations any more than speech analysis is the study of internal phonology. They're different subjects.

You may go on if you like, but you see the point. Though one can set up D, as people do when they do model theoretic semantics or discourse representation or other kinds of theories of semantic values, but it has to be justified on the same terms as setting up phonetic values – you have to show either that it's enabling you to build up an account of things like anaphora (say, along the lines I suggested), or that it's carrying you towards a more transparent picture (a true interface relation), and that's certainly not easy, in fact it doesn't seem possible at all. Notice that there might be an occasional resemblance between the relation R and some relation between words and things in the world. That could perfectly well happen, but that would be an irrelevant accident. And notice again, of course, that we have no intuitions whatsoever about R and D, any more than we have intuitions about phonetic features and their relationships to phonological representations. We can't have intuitions about such things.

Suppose somebody succeeds in justifying this theory of R and D. It's not so trivial. People who work on things like model theory and so on describe what they're doing as semantics (which is very curious, because it has nothing to do with

semantics) and they assume that it's justified. It's kind of funny. No phonologist would get away with that. If a phonologist simply stipulated some set of things and called them phonetic values, and stipulated some relation between them, and then went on to build enormous theories on this, people would laugh. There's no point even looking at a theory of phonetic values unless you give some reason to show that it's an interface, which is what people like Roman Jakobson were trying to do. That's why Jakobson spent a lot of time trying to show that what he thought were the distinctive features had uniform acoustic interpretations and uniform articulatory interpretations and so on, and maybe he is right or wrong, but that's what you have to do. If he had just said, "I like these features and I can make formal theories about them", people wouldn't have even laughed.

The question is, why can you do it on the conceptual side of syntax if you can't do it on the perceptual-articulatory side of syntax? The answer is that you shouldn't. You can kind of get away with it more easily at the conceptual side because we don't know much about the interface.

We have this language faculty over here as part of the mind, and it appears that it has two interface relations, and that just seems like a fact. It seems to interface with the perceptual-articulatory systems and the conceptual-intentional (C-I) systems. That seems to be the two places where the language faculty interacts with other parts of the mind. At the perceptual-articulatory interface, a lot is known. You know a lot about the motions of the articulatory musculature and about sound waves and that kind of stuff. At the conceptual-intentional interface, basically nothing is known. It's mostly mysterious. That's not to say that we don't have evidence about this the C-I interface. In fact, there's plenty of reliable evidence about what semantic relations there are between expressions – relations of entailment and so on – but you don't have any theoretical understanding of the thing at the other side of the interface. Therefore, it's kind of easier to fall into mistakes.

And it seems to me that all of this stuff may not be a mistake, but it has to be justified. To think that it doesn't have to be justified is a mistake. It may be that people made the right guess, but it requires a justification. Short of that, it's on a par with the theory of distinctive features that somebody picked out of a hat. And when you look at the cases, it's not clear that there's going to be an easy answer. The way cases seem to work, it appears that theory of anaphora works by structural properties, not by true semantic properties. That's why things like 'the average man' and 'the young man' come out alike – semantically they're radically different, but syntactically they're alike, and that seems to be what's relevant for a theory of anaphora. This seems to run across the board, as far as I know. Everything in natural

language just seems to care about the form, it doesn't seem to care what the interface relations are, which suggests that we're still working internal to the system. We haven't yet gotten to whatever the interface must be, and it's not in the least clear that any of this is getting near to it.

We're still looking at Frege's notion. Suppose that the internalist postulation of R and D is justified. Suppose that setting up a domain D of syntactic objects and a relation R with reference-like properties to the expressions is justified, meaning there's some way of interpreting the things in D (semantic values) in the world, or at least in the conceptual system and maybe from there to the world – somewhere outside the language faculty. Notice that even if that's true, it's not going to give us any justification at all for believing that there's another relationship R' that holds between expressions and things in the world (or whatever you'd like to say language refers to). Even accepting R and D and having offered the proponent of R and D the conclusion that you can somehow relate D to the world doesn't tell us that there's a relation with the properties of reference holding between expressions and things in the world (this stuff here that we're granting is going to be related to whatever you think language is about – things in the world, things that are believed to be in the world, pick your story). Even if we grant that, it doesn't say that there's going to be any reference-like relation between phrases in the language and those things. It all depends on how this stuff works out, and we haven't a ghost of an idea about that. So, we seem to have no reason at all for believing that there exists a relation of reference that holds between expressions and anything in the world. Nor do we have the slightest reason from any of this to believe in some notion of public language.

These comments are not novel, incidentally. In fact, it was standardly assumed about 40 years ago that it makes no sense to suppose that words have a relation to things. Writing in his introductory logic book 40 years ago, Peter Strawson said that we have to avoid the fallacy of believing in logically proper names, meaning names in natural language that mean things. That's an obvious fallacy, and these are all indications of why it appears to be an obvious fallacy. Unfortunately, the understanding that was pretty standard at that time has mostly been lost, and we now have new approaches coming in which are making exactly the kind of moves that were correctly criticized, in my opinion, by Strawson, Ryle, Wittgenstein, and others back in the early days of ordinary language philosophy. It's kind of striking that if you read the contemporary literature, you almost never find a reference to the fact that these notions were criticized at one time. I mean, there's no refutation of the criticisms. They seem to have been just forgotten. But the criticisms are pretty reasonable, just as the 18th century criticisms that made essentially the same points

were pretty reasonable. I don't think they can simply be dismissed. In fact, they seem correct.

Let's take a closer look at some actual words and see whether it would make sense to pursue the quest for some relationship between words and things. Take, for example, a simple word like 'book'. Some of the things in the world are books, some aren't books, so that looks like a candidate for the term that might refer to things. The word 'book' has some curious properties. For one thing, a book can be concrete or it can be abstract. If I say "the book weighs 5 pounds", it's concrete, because only concrete things weigh. If I say "John is writing a book in his head", it's abstract, or to make it definite, "the book that John wrote in his head" is abstract. In fact, "the book that John wrote" is abstract even if it was published. So, the term 'book' can freely refer either to an abstract or a concrete thing. However, a single occurrence of it can simultaneously refer both to the abstract version and the concrete version. So, if I say, "the book which weighs 5 pounds was written by John". The thing that weighs 5 pounds has no relationship to John at all, he may not even know it exists, but the phrase 'the book' (called a referential phrase) is referring to something which is simultaneously abstract and concrete, and nothing in the world can be simultaneously abstract and concrete, so there can be nothing in the world that is the denotatum of 'book'.

Actually, this is not an accident about 'book', virtually most words are like that. It's the same with definite noun phrases. Take "the deck of cards which is missing a queen is too worn to use". The matrix clause 'the deck of cards is too worn to use' is obviously some kind of physical thing, because that's the only kind of thing that can be worn out. On the other hand, 'the deck of cards which is missing a queen' is some abstract object – in fact, it's some kind of set. We might ask what kind of a concrete object 'a deck of cards that's too worn to use' is. Notice that it's not what's technically called a 'mereological sum' – it's not an entity put together out of its parts. Absolutely not. It can't be that. The left corner of one of the cards does not have the same status in the deck as the card itself, but if it was a sum then that would be true. So, it's not a sum in the sense of the part-whole calculus. It's not a set, obviously. In fact, there's nothing intelligible that it is. Nevertheless, we use it to refer to things, and we use it both concretely and abstractly. Notice that there's nothing paradoxical in terms of the actual notion of referring. Just as I can refer to London without even mentioning it — I can say that he was referring to London when he complained about the traffic. Similarly, I can be referring to a deck of cards when I use the phrase 'the deck of cards', but that doesn't mean that there's a thing in the world – the deck of cards – referred to by the phrase 'the deck of cards.'

What it means is that terms like ‘book’ or ‘deck of cards’, in effect, provide a certain perspective for humans with which to talk about the world, whatever you think the world is. It provides a kind of a lens or a prism or something, ‘Here’s the way to look at things, through this odd perspective’. It’s kind of like Frege’s telescope, if you want to go back to the imagery of *Sinn and Bedeutung*, but the telescope has very curious properties, way beyond what he imagined. And I think one should look at all words that way. All words of any language are not objects that stand in a referential to the world, they’re objects that offer a certain framework or a perspective or whatever for doing what’s called ‘talking about the world’. They may roughly pick out a certain shifting region in a very complex space of human interests and concerns, but they never get any closer to reference than that, except just by accident.

Take the word ‘house’ and consider what you know about houses. Take “John painted the house brown”. What that means is John painted the outer surface brown. If John painted the inner surface brown, he didn’t paint the house brown. So, somehow, the house is its outer surface. The outer surface of the house is some kind of mathematical object, so the house is some kind of abstract object. Similarly, if I say “John saw the house”, it means he saw the outer surface. If John is outside and he’s looking into the window and let’s say there’s a shield around it so that all he can do is look into the window and can’t see anything else, he sees the inner walls fine, but he is still not seeing the house – he has to somehow see the outer surface in order to see the house. So, a house is apparently an abstract object, namely, an outer surface.

On the other hand, it’s certainly not just the abstract object. For example, if a house is just it’s surface, and if John and Mary are equidistant from the surface, then they would be equidistant from the house – and that’s obviously not true because Mary is in the house (assuming). So, if you’re in the house, you can’t be at a certain distance from it. So, the house is not just a surface. John might be near the house, depending on our current conditions for nearness, but Mary can’t be near the house because she’s in it, even though she may be the same distance as John from the house’s outer surface.

If the house is made of wood, then it’s obviously concrete, so we’re not talking about its surface. If I talk about a brown wooden house, I’m referring simultaneously to the surface and its constitution. I can make things worse if I clean the house – I may not touch the outer surface at all, just move things around the inside. So, a house appears to be an exterior surface plus the interior viewed quite abstractly. If I move my house from New York to Boston (I put it in a truck and moved it), a physical

object has been moved, so, it's plainly concrete. So, it's concrete, abstract, it's an outer surface, it's an interior. It has a very strange collection of properties. Certainly, no object in the world can have that collection of properties. It's not that when we refer to houses, we have confused beliefs, and it's not that houses don't exist – the house I live in surely exists. The point is, a house is not a thing in the world with the identifying conditions of 'house', because 'house' doesn't provide identifying conditions for things in the world, it provides a certain way to look at the world.

Actually, all container words are like this. Take 'airplane'. If I'm inside an airplane and say I see the airplane, that could only be true if I see the surface of the wing, or if there's a mirror outside and I see the surface of the airplane reflected in it, otherwise it can't be true. And the rest of the story goes through the same way. It works for 'igloo', it works for things we make up, it works for things that can't exist. If I paint a spherical cube brown, then I painted its outer surface brown, even though it can't exist.

What's more, it doesn't only hold for container words, it holds for terms generally. So, take, say, a mountain, and suppose there's a cave inside it with a light shining in it so I can see, and there's a way to look in. I'm out here and I'm like John trying to look into the house. I can see what's in the cave fine, but I'm not seeing the mountain, even though 'mountain' isn't a container word. If that's true then 'mountain' is again something that's pretty abstract. The very thing is a 'mountain' or an 'island' or a 'plateau' depending on other things, which don't affect the thing at all.

Again, the words that are used, at most, offer a certain quite complex perspective for viewing the world. You might think this doesn't work for proper names at least, but, unfortunately, it works for them as well. So, take the standard proper names that appear in the various puzzles and paradoxes. Take 'London'. Certainly, London exists. It's not a fiction. We agree on that. It's not a town in some fairy story. But what is it? Suppose I say, "London is ugly". Presumably, I'm referring to the buildings or something or other. Suppose I say, "London has gotten to be a very unhappy place during the Thatcher years" or "London is taking on a much harsher edge during the Thatcher years", then I'm referring to the people and the way they live. If all the people leave London (say, in an earthquake or something), it's still London, even though when I talk about London as being an unhappy place, I'm only referring to people. If I say, "London is polluted", I'm talking about the air above London, though not too high. Suppose I put those things together, "London is ugly, unhappy, and polluted and it should be destroyed and rebuilt a hundred miles up the Thames". That all makes sense, but what kind of thing

is London now? – it's a thing that involves only people, that involves only the air above it, that involves only its buildings, and that can be destroyed and built somewhere else and still be London. Surely, nothing in the world can have those properties. And if we go further, we get even stranger things.

Now, if you believe that there's a common language called English in which the word 'London' has referential properties – it picks out a thing – you're naturally going to get into tremendous paradox, because that belief is so exotic to start with that it's going to quickly lead to paradox. And, in fact, one famous paradox is Kripke's puzzle of belief, in which he shows, correctly, that if you believe all these things (which he takes for granted), you get into the problem of somebody believing that London is pretty and London is ugly without being irrational. I won't run through the example, but his logic is perfect. That's the conclusion you get to, and you get to it from the assumption that there is a common language English with the word 'London' which has a referential relation to some thing, that is, you get into the paradox of the puzzle of belief by basically assuming the Fregean picture. If you start by taking some weird properties of 'London' and thinking that it's going to pick out a thing in the world, of course you will get into all sorts of paradoxes.

On the other hand, if you think of 'London' as being a city name and as offering a kind of perspective for talking about the world that this category of linguistic expressions does, then there's no paradox. Kripke's paradox happens to involve somebody who's bilingual and knows both 'London' and 'Londres' (which he says pick out the same thing), and those two concepts could just fit differently into a person's belief system without paradox at all.

We can go on and give more examples and all of them would show the same basic thing. It's just a mistake to believe that there's a reference-like relation between words and things. That's not the way language works. The way language works, apparently, is that it offers certain conceptual perspectives for talking about things. That's all, there's no closer relationship. Maybe we'll want to set up technical notions like R and D for internalist reasons, but there doesn't seem to be anything beyond that.

I might say just as an aside that a lot of the metaphysical conclusions that are drawn from these assumptions also seem dubious. So, take, say, Kripke's discussion in *Naming and Necessity* which led to the contemporary essentialist literature. He says things which are extremely plausible. In fact, they seem certainly true. But he draws conclusions from them which are startling, and he intends to be startling, and

it's not at all clear that they should follow – they follow via the assumption that proper names have reference relations to things in the world.

So, take one of his cases (this is written in 1972 so he's talking about Nixon), "Nixon might not have been elected in 1968". So, it's perfectly sensible to say, "Nixon might not have been elected in 1968 if only the Democrats had played the game right". But if you say "Nixon might not have had his own parents", that sounds impossible. If Nixon had had different parents, he wouldn't be Nixon. On the other hand, if Nixon hadn't been elected in 1968, he'd still be Nixon. There's a clear distinction and I don't think anybody has a problem with it. And Kripke therefore draws the conclusion that Nixon has certain essential properties – like having certain parents – and other accidental properties – like being elected in 1968. It's not a demonstrative argument but that's certainly plausible if you think that Nixon is a logically proper name – a name that picks out a certain thing in the world (namely, some object), through the relation of reference. On the other hand, if you think that the word 'Nixon' works like 'London', then those conclusions don't follow at all. 'Nixon', after all, is not a logically proper name, namely, a name divorced of all properties other than its referential properties. 'Nixon' is a personal name, which is a category in language, just like city names are categories in language. Whether personal names are conceptual or linguistic, we're not sure, but they're certainly a category. So, when I talk about 'Nixon', I'm really saying 'the person Nixon', and it's certainly true that 'the person Nixon' couldn't have had different parents, because you're not the same person if you have different parents, but that's a de dicto necessity – something you draw from the words – it has nothing to do with the facts. It's saying that persons are the kinds of things that are identified in terms of their parents. That's a logical necessity. It has nothing to do with the world. So, it doesn't tell you anything about the essential properties of Nixon.

Suppose we try to make up a concept of Nixon which is a logically proper name. That's a hard thing to do. We're now off in some branch of mathematics because language doesn't have things like that. But let's try. Let's invent 'N' as a logically proper name referring to Nixon and has no properties other than that it refers to Nixon. Now, at least, my intuitions are gone on everything. I certainly don't have these intuitions. In fact, now I would think Nixon is a different entity just if he combs his hair differently. Just viewed as an entity (and not a person), if he combs his hair differently, he would be a different entity. A Martian who categorizes things in terms of how hair is combed would draw exactly that conclusion. In fact, if the Martian was confused into thinking that his name 'Nixon' was a logically proper name (where this wouldn't be persons anymore but whatever category distinguishes one type of hair combing from another), he might conclude that one of Nixon's

essential features is that his hair is combed a particular way – that's a fact about the world – and he would be just as justified in saying that as Kripke is in saying that having particular parents is an essential property of Nixon.

All of these startling conclusions follow from the first move, which is to accept Frege's framework which claims there is a relation R holding between words and things. If you accept that, you're off and running and all these things follow and you get to a lot of very strange conclusions, none of which are necessary and none of which even seem true, because there doesn't seem to be any relation like R.

Does any of this have anything to do with Frege's project? That's not so clear. It's not at all clear that Frege cared about natural language or natural thinking. It's true that he writes about it, but if you look, there's a footnote here and there which says that ordinary languages aren't perfect languages, so they have various problems, we must fix them up, and so on. It looks as if what he was actually talking about, if you think through his work, is what he called a 'perfect language', a language which doesn't have the imperfections of natural language, i.e., a created symbolic system which won't have the properties of language but will have perfect properties – it will be perfect for his purposes. And his purposes were basically metamathematics (as it's now called) – a language for talking about mathematics. And for those purposes it might make perfect sense to assume those three conditions – there's a common store of thoughts, there's a common public language in which they're expressed, and there's a relation of reference that holds between the elements and the things. That may make perfect sense for arithmetic.

If you want to talk about arithmetic and math generally, you have things like '16' and 'e' and so on, and you want names that will refer to those things, and you want variables which will range over those things. You don't want those names and those variables to have any other properties, and they're going to be logically perfect. You want this to be a public language, everybody could use it whether they're Japanese or African or whatever. And you want the thoughts expressed in it (if that's an appropriate notion) to be a public store. So, a guy who uses it in Japan will have the same thought I have when I use it.

All of those are perfectly reasonable objectives for the project of making up a language for mathematics. Insofar as that's true and insofar as we take that to be Frege's project, everything I say is irrelevant to Frege's project. But it's not irrelevant to everything that's been drawn from Frege's project, which is the application of it to natural language and natural thought.

Now, whether Frege intended that or not is a textual question. Actually, it doesn't seem that he made a very clear distinction, but maybe he did, maybe he didn't, that's not really the point. The point is that we should make a clear distinction between studying things in the natural world (like language and thinking) and inventing symbolic systems for particular purposes (like metamathematics or (probably) science). It seems to me reasonable to claim that part of the project of the natural sciences is to aim for a Fregean style symbolic system, i.e., a symbolic system that will have a common store of thoughts, a public symbolic system in which to express them, and a reference like relation, meaning that you want the terms of that system (whether names or variables) to denote real things (like electrons and stuff), and you'll shift around your system until you get that property. If you find that you're not picking out real things, you'll switch the system around, because your project is to create a Fregean style perfect language.

Now that language differs from natural language in crucial respects. For one thing, it differs totally in syntax – it will use calculus and all kinds of things, and it will throw out most of the syntax of natural language, which is a mess for these purposes.

It differs at the semantic level in two crucial respects. For one thing, it will never use terms with the weird properties of natural language terms – terms that offer ways of talking about things that are both abstract and concrete. In fact, it's hard to imagine that any term of natural language would even survive into that system. Basically, that's what happens – as science advances, all the terms go. They may use the same sound but with a different meaning, so it's a different word. The symbolic system will deviate from natural language, in that it will never use terms with properties of natural language terms. And it will deviate in the second respect, in that it will add semantic properties that natural language doesn't have – like reference. And it, probably, won't need modes of fixing reference, because the fixing of reference is done by stipulation, so you don't have to worry about Sinn.

In any event, that might be quite a reasonable project for a symbolic system for science and for mathematics. In fact, it would be a perfectly reasonable project for the theory of natural language, because after all the theory of natural language is a scientific theory, so it will be following the scientific project. But, of course, it would be a vulgar error to invest the theory of X with the properties of X – you must keep them separate, of course. So, natural language is what it is, and we may pick a symbolic system to talk about it, which doesn't have natural language properties. In fact, obviously, that's what will happen.

This is a huge topic and I've barely skimmed the surface of it. As far as I can see, there is no reasonable alternative to a completely naturalistic approach to language. Everything that looks like an alternative seems to me to collapse. However, I don't want to leave you with the misimpression that I'm saying that a naturalistic theory of language can answer the main questions about language. In fact, it seems that it can't. Like, the question of the creative aspect of language use, which, traditionally, was the core property of language, the very criterion for the existence of other minds, that isn't even approached by a naturalistic theory of language. Why? Maybe the theory is wrong, maybe we're not smart enough, maybe something else, but one strong possibility is that those questions are simply not within our cognitive reach. Those are mysteries for us, not problems. There's nothing dualist about that. It's just a comment about a thing in the world, namely, us and our cognitive character. There is no dualism at all, even if the domain of mysteries would happen to correspond closely to one of the traditional domains of metaphysical dualism.

Now, if this whole discussion is at all on the right track, one would want to know why non-naturalistic dualist positions have such a strong appeal to the imagination. What I've been suggesting for the last five days is that they come up all over the place, even on the part of people who regard themselves as real hard-headed scientists. If you look closely, I think you find that they're mired in non-naturalistic dualism, and it seems to cover just about every way of looking at the field you can think of. If what I said is correct, almost everything is some form of non-naturalistic dualism, and that seems strange, because nobody accepts dualism. In fact, everybody ridicules it.

So how come (assuming this is correct) everybody is caught up in it? Here you have to speculate, and the plausible speculation seems to be the one that I mentioned earlier – our common-sense view of people is irremediably dualist. We see people as a combination of a body and a soul. It's just the way we see them. That's why we, basically, define 'person' in terms of psychic persistence – you are the same person if you have the same psychic pattern, not if you have the same physical pattern – because that's how we look at people. We look at people in terms of a body and a soul, and we can't help it any more than we can help seeing the sun setting. We know it's not the right story, but we can't see it any other way, and it's possible that we just can't see people other than in dualist terms.

However, if that's correct, we should certainly not allow that to hamper our efforts to discover what kind of creatures persons are, any more than we let these perceptual constraints hamper our inquiry into what the world is.

So, it could be that this is right (and I sort of suspect that it probably is), but then it will just have to go the way of the mechanical philosophy and everything else that's been thrown out that was self-evident. The ways in which we look at the world are just not correct – they don't give you a correct picture of the world. You have to turn to the science-forming capacity and struggle to approach the right way of looking at the world in some other fashion, throwing out your intuitive conceptions because they're not relevant. That's the basic lesson of Newtonian physics, I think.

THE END

Transcribed by: Nishant Mulay (nishantmulay@gmail.com)