Abstract and Keywords

Human language is not the only naturally occurring symbol system. There are many animals other than human beings that communicate by means of signs or signals; vervet monkeys, for example, have specialized warning cries for different kinds of predators. And some animal-communication systems even have a rudimentary syntax: the dances performed by certain honey bees have structural elements that tell other bees the direction and distance from the hive of a nectar source. But what’s distinctive of human language — and the feature that Descartes was highlighting — is that the syntax of human language permits us to take parts of signs and recombine them with parts of other signs.

Keywords: symbol system, human language, animal-communication systems, reflex behaviour, human syntax

Preliminary Thoughts about Thinking

Cogito, ergo sum, Descartes reasoned, and then immediately asked, ‘What is this “I” that thinks?’. This essay asks the different question, ‘What is this thinking that I do?’.

Perhaps the first thing to be said about thinking is that it is a mental activity. But what does that mean? Most people, in speaking of the ‘mental’, have in mind some kind of contrast with the physical. Physical activities all involve our bodies, or parts of them, in some conspicuous way. Physical actions that we perform—walking, writing, playing a musical instrument—all involve the voluntary movement of our limbs. Physical processes like perception and respiration are clearly dependent upon the well functioning of our bodily organs. Thinking seems different from both those things. I can think without
Thinking

moving any part of my body; I can think with my eyes closed, and my ears plugged. The only thing that seems to be necessary in order for me to think is that I be conscious. Could it be that thinking doesn't require a body at all?

René Descartes, notoriously, came to precisely this conclusion. In Meditations on a First Philosophy he argued that because mind and body were conceptually distinct it was possible for the one to exist without the other. Hence, mind and body were actually distinct, even when they were co-instantiated (Descartes 1641/1901). This is an argument for dualism that many philosophers still find compelling today. But there's a second, less-well-known argument of Descartes's for the same conclusion. In this argument, from his earlier work the Discourse on Method, Descartes appeals to some very specific—and plausible—assumptions about the nature of thought. Even if (as I believe) the argument fails to establish the immateriality of thought, it makes a good starting point for our investigation.

The argument comes as Descartes is extolling the marvels of animal anatomy. He acknowledges that the complex structure and finely tuned operations of biological organs make it seem as if living beings might be machines of a sort. While he thinks that this suggestion is perfectly apt in the case of non-human animals, he insists that it is completely wrong when it comes to human beings. A machine, he says, designed with the internal workings and external appearance of an animal, would be indistinguishable from the real thing. But no mechanical simulacrum of a human being, no matter how perfectly crafted to resemble us outwardly, could pass for a real person. There would be, he says, ‘two most certain tests whereby to know that [such machines] were not therefore really men’. The argument for dualism comes in the second of these two tests:

[A]lthough these machines might do several things as well or perhaps better than we do, they are inevitably lacking in some other, through which we discover that they act, not by knowledge, but only by the arrangement [disposition] of their organs. For, whereas reason is a universal instrument which can serve in all sorts of encounters, these organs need some particular arrangement for each particular action. As a result of that, it is morally impossible that there is in a machine’s organs sufficient variety to act in all the events of our lives in the same way that our reason empowers us to act

(Descartes 1637: pt. v, para. 35)
To understand the argument, we must first appreciate the distinction that Descartes is drawing between two ways in which seemingly intelligent behaviour might be produced: it might be simply necessitated by the machine's antecedent physical state—by the 'disposition of its organs'—or it might be produced 'by knowledge'. Only in the second case, Descartes says, is there genuine intelligence at work, for it is only in this case that the appropriateness of the behaviour to its circumstances is due to an understanding of the situation. In the first case the appearance of intelligence behind the behaviour can be explained by the fact that the physical structure that produces it happens to be (or, in the case of a machine, was designed to be) specialized to produce precisely that kind of behaviour, in precisely those circumstances. We can be confident that this is the explanation, Descartes argues, because the apparently skilled behaviour displayed by animals or machines is always limited to one or a few domains. In contrast, Descartes says, our faculty of reason, which enables us to act from knowledge, is 'a universal instrument which can serve in all sorts of encounters'.

Let's suppose for the moment that Descartes has established a categorical difference between two types of behaviour production. (Let's call the first type 'brute causation' and the second type 'rational causation'.) How does he get from this premise to the conclusion that rational causation cannot transpire within a material medium? He says of brutes' 'souls' as compared with human beings' that 'once one knows how different they are, one understands much better the reasons which prove that the nature of our souls is totally independent of the body'. How does the rest of the argument go? We need to understand better what Descartes takes to be involved when behaviour is 'due to knowledge'.

In order for us even to suspect that a piece of behaviour is genuinely intelligent, the behaviour has to be appropriate to the circumstances in which it is displayed. Contrast, for example, the following two cases: in the first, my doctor strikes my knee with her little mallet and my foot kicks out; in the second, I kick my foot and hit an assailant who has just struck me with a mallet (and seems about to do so again). In the first case there is nothing that makes my kicking appropriate to the circumstance of having been struck by a mallet; it is simply what my foot does when my knee has been struck in that way. In the second case, however, it makes sense for me to kick, because doing so may well drive away my assailant. The first case is a clear example of reflex action—hardly an 'action' at all—and the second a clear example of rational action. What's the difference?

In the second case, as we know, Descartes would say that my behavioural response to the circumstances I face is mediated by knowledge. That means that one of the antecedents of my behaviour is a representation of the situation I face. In fact, there are several representations involved. I need to represent not only my circumstances, but the goal I wish to achieve, and the means by which I might achieve it. These representations stand in rational relations to each other. Because I want to drive away my assailant, and because I believe that kicking her is apt to drive her away, I conclude that kicking her would be a good thing to do. If these representations are all accurate—if they correctly represent my situation, my goals, and my options—then the behaviour that results from this process will be appropriate to the situation: it will further my goal.
Here, then, are two surpassingly important features of thoughts: First of all, they represent—they display what the Austrian philosopher Franz Brentano called 'intentionality': the feature of standing for, or being about, something else (Brentano 1874/1973). Second, they possess some kind of rational structure in virtue of which they can stand in logical or rational relations to each other. It is likely that Descartes had both these features well in mind when he spoke of actions being due to knowledge.

Returning to the first of my two illustrative cases, the case of reflex ‘action’, we can now see that the arbitrariness of the kicking response relative to the circumstance of being struck by the hammer is due to the fact that the kicking response is not in any way sensitive to my representation of my situation. We can better appreciate the difference by considering counterfactuals in the two cases. In the second case, if I had not represented the situation as one in which I was being assaulted then I would not have kicked. But in the first case the ‘disposition of my organs’ would have determined me to kick regardless of how I viewed the situation. It is this feature of rationally caused behaviour that makes for flexibility. In so far as my thoughts accurately represent my situation, goals, etc., and in so far as my behaviour is counterfactually dependent on my thoughts, my behaviour will tend to be appropriate to my circumstances whatever they are.

This, then, is the categorical distinction Descartes needs: between beings that can and beings that cannot condition their behaviour to representations of their circumstances, goals, and options. But why think that non-human animals and machines fall into the second category rather than the first?

Reflex behaviour, of course, need not be inappropriate to circumstances; it may mimic intelligent behaviour. This is the case, Descartes believes, with the behaviour of animals and machines. The instinctive behaviour of animals or the operation of well-designed machines may even be more efficacious than human efforts—a person who tried to build a nest, for example, would be no competition for a bird, and our intuitive sense of the time is no match for a clock's marking of the minutes—but the true aetiology of such behaviour would be betrayed once we placed the animal or the machine in a novel circumstance. Because the brute or the mechanism does not know what is going on—because it lacks the capacity for representation—it cannot tailor its behaviour to suit its circumstances. Indeed, Descartes argues, it is the fact that animals’ and machines’ performance is so uneven—better than human behaviour in some circumstances, but worse in most others—that assures us that they are not acting through reason. If the better performances were due to thinking, then we would expect the performer to be uniformly better than us, able to produce more appropriate behaviour across the board:

Thus, the fact that they do better than we do does not prove that they have a mind, for, if that were the case, they would have more of it than any of us and would do better in all other things; it rather shows that they have no reason at all,
and that it’s nature which has activated them according to the arrangement of their organs—just as one sees that a clock, which is composed only of wheels and springs, can keep track of the hours and measure time more accurately than we can, for all our care.

(Descartes 1637: pt. v, para. 38)

So the argument for dualism proceeds: since beings that lack reason can only produce appropriate behaviour through the operation of specialized organs, they would need a separate specialized organ for every distinct task they face. It would be impossible—morally impossible, Descartes says, and this is important—to outfit a purely material being with enough of these organs to enable the being to respond appropriately to any circumstance whatsoever. Since we human beings do possess this kind of general flexibility, the faculty that enables it—‘reason’, in Descartes's terms—must inhere in an immaterial substance. But there’s still a piece missing: why is Descartes so confident that we possess the degree of flexibility that would require an impossibly large inventory of specialized organs in a purely material being? What makes him think that reason is truly a ‘universal instrument’? The answer comes in the first test he gives for telling whether one is dealing with a person or a mechanical simulacrum, and has to do with language. Machines, Descartes says,

would never be able to use words or other signs ... as we do to declare our thoughts to others. For one can easily imagine a machine made in such a way that it expresses words, even that it expresses some words relevant to some physical actions which bring about some change in its organs ... but one cannot imagine a machine that arranges words in various ways to reply to the sense of everything said in its presence, as the most stupid human beings are capable of doing.

(1637: pt. v, para. 34)

Notice that when Descartes asserts that no machine could use language, he is very specific about the limitation he has in mind. He does not doubt that one could build a machine that would produce what sound like meaningful utterances, and produce them in circumstances for which they appear appropriate—a machine, for example, that produces the sound ‘Please don't do that’ if we strike it. The thing Descartes thinks a machine could not do is to ‘arrange words in various ways to reply to the sense of everything said in its presence’. That is, no machine can recombine the parts of its ‘utterance’ so as to compose new utterances that make sense relative to new circumstances. Descartes is here pointing to a strikingly distinctive feature of the human communication system: its compositionality. We've already remarked that thought is representational, that it has intentional content, or meaning. Language does, too: it is composed of symbols that stand for other things. The set of relations between symbols and the things they represent is called the semantics of the symbol system. While all symbol systems have semantics, only some have syntax—rules that govern the ways in which primitive signs can be combined so as to produce more complex signs.
Human language is not the only naturally occurring symbol system. There are many animals other than human beings that communicate by means of signs or signals; vervet monkeys, for example, have specialized warning cries for different kinds of predators. And some animal-communication systems even have a rudimentary syntax: the dances performed by certain honey bees have structural elements that tell other bees the direction and distance from the hive of a nectar source. But what's distinctive of human language—and the feature that Descartes was highlighting—is that the syntax of human language permits us to take parts of signs and recombine them with parts of other signs. The vervet monkey can utter a shriek that tells its troupe that an eagle is approaching overhead, and a different kind of shriek that tells them that there's a snake nearby in the grass. But each of these signs is an indissoluble unit—the monkey cannot recombine parts of the two shrieks to express, for example, the thought that there's an eagle in the grass (Cheney and Seyfarth 1990). In contrast, any human being who can say both that there's an eagle overhead and that there's a snake in the grass can also say that there's a snake overhead or that there's an eagle in the grass—this is guaranteed by the way sentences of human languages are composed. Jerry Fodor and Zenon Pylyshyn call this property systematicity: your communication code is systematic if in possessing the ability to express one thought you are automatically equipped to express many (Fodor and Pylyshyn 1988).

But there's another important property of human syntax: it is recursive. This means that we can take the expression that results from the application of some rule of combination and reapply that same rule to the new expression. For example, I can form a complex noun phrase by conjoining two nouns: 'Bob' and 'Carol' → 'Bob and Carol'. But I can also reapply that rule to the noun phrase I just constructed: 'Bob and Carol' and 'Ted' → 'Bob and Carol and Ted'. Compositionality plus recursion give us a guarantee that language has the resources to generate an unlimited number of distinct, meaningful signs. This property is called by Chomsky creativity (1975) and by Fodor and Pylyshyn productivity (1988).

Descartes clearly took the compositionality of language to be evidence that human thought was itself systematic and productive. Because language is compositional and recursive, we know that we are not like the vervet monkeys, stuck with a fixed repertoire of signs. But if we have an unlimited number of distinct linguistic signs available to us, related in a systematic way, then we must also have available to us an elastic repertoire of distinct thoughts, corresponding to each of these signs. Hence, our observed ability to utilize a compositional (and recursive) language is ipso facto evidence of our ability to think and express an unlimited number of distinct thoughts. And that's the premise that Descartes needed in order to complete his empirical argument for the immateriality of thought.

But here lies the great irony. In highlighting the systematicity of thought, Descartes inadvertently laid the groundwork for a materialist theory of mind. If it is the mind's systematicity—rather than, say, its subjectivity—that is the sticking point for a materialist
account of mentality, then the *Discourse* argument for dualism could be neutralized if someone could show how systematicity could be mechanized. And this, it turns out, is exactly what Alan Turing did, more on this soon.

As we’ve seen, Descartes presumed that the structure of language mirrored the structure of thought—that language was, indeed, merely the outward expression of thought. The most important aspect of this isomorphism was rational structure, secured by compositionality. This emphasis on rational structure was one of the things that distinguished philosophical rationalism from the empiricism that emerged in Great Britain about half a century after the publication of Descartes’s *Discourse* (Locke 1690/1975). So let’s be more explicit about what compositionality involves. We’ll necessarily go beyond Descartes’s brief remarks in the *Discourse*.

Language is compositional: its primitive parts are words, which can be combined to create phrases and sentences. Does thought, similarly, have parts? It appears that it does. Consider the thought that HERSHEY’S KISSES ARE MADE OF CHOCOLATE.2 This thought seems to contain, as components, several elements: one that is about Hershey’s kisses, one that is about chocolate, and one that is about being made of something. In this respect, thought contrasts interestingly with another kind of mental state, a ‘qualitative’ state like the one I would experience were I to taste a Hershey’s kiss. The experience of tasting the candy doesn’t seem at all to be segmented in the way the thought is—I would be hard pressed to distinguish a HERSHEY’S KISS part of a candy-eating experience from a CHOCOLATE part.3

We have a name for thought parts: we call them ‘concepts’, and we individuate thoughts according to the concepts that compose them. The thought that lobster is high in cholesterol is a different thought from the thought that the most expensive thing on the menu is high in cholesterol, because the concept LOBSTER is different from the concept THE MOST EXPENSIVE THING ON THE MENU even if, in a given circumstance, the most expensive thing on the menu happens to be lobster. In other words, it’s one thing to think about lobster as lobster, and another thing to think about it as the most expensive thing on the menu. Thoughts enable us to do this, and they do so by permitting us to pick out the same thing under different concepts.

This feature of thought is called ‘intensionality’ (and sometimes called ‘intensionality-with-an-s’ to distinguish it from ‘intentionality’—‘intentionality-with-a-t’).4 The intensionality of thought explains some otherwise puzzling phenomena. Lois Lane, as any well-educated person knows, believes that Superman is strong and brave. She does not believe that Clark Kent is strong and brave—quite the contrary. She believes that Clark Kent is weak and cowardly. But it turns out that Superman is Clark Kent—he disguises himself with glasses and hairstyling! What, then, should we say about Lois’s state of mind? We might insist that Lois does believe that Clark Kent is strong and brave—quite the contrary. She believes that Clark Kent is weak and cowardly. But this is something Lois would vehemently deny. Maybe we should say that Lois believes contradictory things. But that would indict her of logical confusion, and there’s no reason to think she has that problem. Rather, what seems clear is that Lois thinks of, or conceives
Thinking

of Superman in two different ways: as SUPERMAN, and also as CLARK KENT. Because she has two distinct concepts of the same individual, her SUPERMAN thoughts are distinct thoughts from her CLARK KENT thoughts.

Brentano held that intentionality was the ‘hallmark of the mental’—that what made a state a mental state was its being representational. It is currently a hot topic in philosophy whether this claim is true. It is clear that sensations and emotions are part of our psychologies, but not so clear that they have representational content; some philosophers say they do, and some say they do not. But if such states do represent, they do it in a different way than thoughts do. I do not have available to me different ways of experiencing the taste of the candy, one a ‘Hershey's kiss’ way and one a ‘chocolate’ way. There is one distinctive way that a Hershey's kiss tastes, and that one way defines the qualitative character of the experience. Relatedly, my experience of the taste of a Hershey's kiss does not depend on my having any knowledge or understanding of ‘Hershey-ness’. I need the concept HERSEY'S KISS in order to think the thought ‘A Hershey's kiss is made of chocolate’, but I don't need it to experience its chocolate-y goodness. We can capture these differences by saying that thoughts are conceptual while qualitative mental states are non-conceptual.

Not all concepts are the same. The concepts HERSEY'S KISS and CHOCOLATE are different in a crucial way from the concept IS MADE OF. Roughly, the first two concepts are about things, while the third is about a property of, or a relation among, things. These different kinds of concept combine in specific ways, giving thought its structure. We noted above that concepts can combine and recombine in lots of different ways, provided they are of the right type to work together. To a first approximation, we can say that any concept of the ‘thing’ type can combine with any concept of the ‘property’ type. (In language, the mirror distinction is between subject terms and predicate terms.) This is, obviously, not all there is to conceptual structure, but it may be the fundamental feature of conceptually structured representation (Evans 1982; Fodor 2007).

The version of computationalism that I'll present below is Cartesian or rationalist, in that it takes for granted that thought is structured compositionally. I’ll have a few words to say at the end about the empiricist alternative, and the kind of computational model consonant with it.
The Computer Model of Mind: Thinking as Information Processing

At first glance it would seem that no two things could be more deeply at odds than creativity and mechanism. As we just saw, Descartes relied precisely on the intuition that deterministic processes could not give rise to novelty to frame his second argument for dualism. But important work in logic in the 1930s by logicians Kurt Gödel, Alonso Church, and Alan Turing posed a deep challenge to this intuition. In particular, Turing proved that a deterministic ‘machine’—really a mathematical object, rather than a physical device—could embody recursive procedures (1937). In its essentials, the machine was conceived as consisting of a tape, on which is written some pattern of distinguishable symbols (1s and 0s, for example), and a writer/scanner that can move right or left along the tape, registering and reacting to the symbol it finds. The operation of the machine is specified by a list of ‘instructions’ (the reason for the scare-quotes will come later) laid out in its ‘machine table’. Each instruction specifies what the machine will do, depending on what state the machine is currently in, and on what input the machine receives. The instruction could tell the writer/scanner to move left or right or to stop, and to erase, erase and replace, or ignore a symbol appearing on the tape. The table also specifies a starting state and an ending state for the machine. The configuration of symbols on the tape at the start represents the input to the machine, and the configuration at the end represents the output.

What Turing showed was that for any step-by-step procedure that could be used to calculate the value of a function for a given argument, that procedure (or algorithm) could be captured by some Turing-machine table. Hence, if a function is computable at all—that is, if there is an algorithm guaranteed to produce a value for any argument—there will be a Turing machine that can compute it. Thus, for example, there is at least one Turing machine that embodies an algorithm for addition function, and at least one for subtraction, and so on. Showing this much is an achievement dazzling enough, but Turing also went on to show something utterly breathtaking. He proved that one single machine—a ‘universal Turing machine’—could do anything and everything that could be done by any simple Turing machine. The trick was to show that one could effectively code and then list all the different special-purpose (or ‘simple’) Turing machines; one could then build a Turing machine that took the list positions and coded instructions as part of its input. In this way, a single machine could, in effect, look up the simple machine that embodied the function to be computed, and then just follow the instructions associated with that machine.

The universal Turing machine’s ability to simulate the activity of any simple Turing machine affords it precisely the kind of flexibility that Descartes thought was beyond the capacities of ordinary matter: the flexibility to make apposite responses to an unlimited number of distinct problems. Moreover, the machine’s flexibility is the result of the systematicity of the symbol system that encodes its inputs, outputs, and instructions. The
same two symbols—‘1’ and ‘0’—are combined and recombined in rule-governed ways to represent a potential infinitude of meanings. Yet the machine’s behaviour—no matter how complex—is always decomposable into small step-by-step reactions to primitive elements of the complex symbols.

Turing's mathematical results were quickly applied to physical reality, resulting in the construction of concrete ‘thinking machines’: digital computers. These were not literally Turing machines—they incorporated some architectural innovations due to John von Neumann, for example data stores, or ‘memory’—but still realized the essential feature of Turing’s model: computation performed on structured symbols. The philosophical import of these developments was also rapidly exploited. Hilary Putnam noted that the relation between the machine table of a computer and the ‘hardware’ that implemented it offered a useful model for representing the relation of mind to body (Putnam 1960, 1967a, 1967b). Mental states, on this model, would be abstractly characterized states of a human brain, just as machine states were abstractly characterized states of an electronic mechanism. Because the defining features of these abstract states were the functional relationships among individual states, inputs and outputs, this general view of mentality became known as functionalism.

If we take the machine analogy completely at face value, we get the view of thinking known as the computational model: thinking is computation. That is, a thought process consists in a series of transitions from brain state to brain state, in a way that mirrors an explicit inference or computation, just as the electronic states of a digital computer each embody steps in an inference or computation. Jerry Fodor pressed the analogy a step further. He argued that if thinking was computation, then there needed to be a medium of computation—a symbolic language over which the computational processes that constituted thought were defined. Thoughts, then, would be tokenings of sentences in an internal language. This ‘language of thought’, he argued, was analogous to a computer’s ‘machine language’, the system of symbols to which the machine's most basic computational operations are sensitive. Electronic computers can translate machine language strings back and forth into more 'user-friendly' programming languages by means of special translation programs (compilers). Analogously, we can regard the acquisition of human natural languages as the building of a translation program that allows us to move back and forth between ‘Mentalese’ and ordinary public language (Fodor 1975). We'll return to the idea of a mental ‘machine language’ below.

Before leaving this point, however, let me just note that Fodor's view raises a possibility that Descartes would not have allowed; namely, that ‘dumb brutes’ might be thinkers. Descartes, as we saw, presumed that the ability to use language tracked the ability to think—that all and only language users possessed the faculty of reason. (This is so even though his argument requires only that language use be evidentially sufficient for the attribution of reason to a creature.) But if Fodor is correct, the possession of an internal language of thought—a biologically specified ‘machine language’—is independent of the capacity for speech. We can thus make sense of attributions of genuine thought to non-linguistic creatures—infrahuman animals, as well as pre-linguistic human children; to say
Thinking

that such creatures think is to say that their behaviour is mediated by an internal compositional system of representation. This is indeed the position taken by contemporary ethologists and developmental psychologists, who have evidence of behavioural flexibility and creativity in animals and infants that might have changed Descartes's assessments (Gallistel 1990; Carey and Spelke 1994).

Although Turing did not develop a computational theory of human thought, he did commit himself explicitly to the proposition that a machine could—at least in principle—think. Ironically (again), Turing endorsed the very same criterion of thinking that Descartes appealed to in his argument for dualism (Turing 1950). Turing, like Descartes, thought that the sustained ability to respond appropriately to a variety of circumstances was sufficient for being a genuine thinker. And like Descartes he saw verbal interaction as providing the crucial test. It's just that, unlike Descartes, Turing thought that this ability could be embodied in a machine.9

Turing envisioned an experimental set-up like the following.10 Position a human ‘interrogator’ at two keyboards, one of which feeds inputs to another human being, and the other of which feeds inputs to a computing machine. Both human being and machine can respond to the inputs, and their responses are displayed as lines of type on a display screen visible to the interrogator. The interrogator does not know which keyboard is connected to which responder, and so does not know (at least initially) which responses are coming from which responder. The interrogator types questions or instructions, like ‘Please write me a sonnet on the subject of the Forth Bridge’ (or, if in the present day, something more like: ‘Who do you think should have custody of Britney's kids?’) and would receive answers like: ‘Count me out on this one; I never could write poetry’ (‘Don’t ask me; I don’t watch television’). If after a sufficient amount of time (Turing leaves this unspecified) the interrogator still cannot tell which of the responders is the machine and which is the human being, then the machine is deemed to be a genuine thinker.

It's noteworthy that Turing's test of intelligence imposes no conditions on the means by which the machine is allowed to accomplish its deception. Most philosophers have taken this to mean that Turing was proposing a behaviourist definition of ‘thinking’: that he was essentially saying that thinking is whatever might turn out to be responsible for intelligent-appearing behaviour. If this is what Turing had in mind, it's clear that his 'definition' is inadequate, for it's easy enough to describe a hypothetical case in which an obviously dumb device manages to pass the Turing test.

To see how this might work, let's consider a hypothetical machine devised by Ned Block (1995). The machine runs a simple program, constructed in the following way. First a giant list is made of all possible strings of symbols that could be typed by a human interrogator within some particular period of time—say an hour. The list will include a vast number of what we might call ‘conversations’: strings that look like transcripts of some sensible conversation between two human beings. The next step, therefore, is for a team of programmers to comb through the list of strings, pulling out every conversation, and throwing away all the other strings, the ones that do not resemble actual human
exchanges. The programmers then ‘punctuate’ each conversation on the list by separating with commas or some other delimiters the lines that would belong to one ‘conversant’ from those belonging to the other. Call the delimited parts of the strings ‘remarks’.

Now the instructions to the machine are trivial. All it has to do is to match strings of remarks with conversations on the list. It can do this step by step. When it receives its first remark as input, it finds some conversation that begins with that remark, and then copies the second remark in that conversation as its output. (Remember that every remark that is intelligible and that can be typed in the space of an hour will appear as part of some—indeed many—conversations on the grand list.) The human interrogator will, of course, make some reasonable response to the machine's remark. The machine now finds a conversation containing the string consisting of the first remark, the machine's own remark, and the interrogator's next remark, and outputs the fourth remark in that conversation. (Again, every possible three-remark conversation will be on the grand list.) This process is repeated until the hour is up.  

Will this work? Many actual working programs operate on essentially these principles. Think of (the now ubiquitous) phone menus. Suppose you want to renew a pharmaceutical prescription. You dial the number of the pharmacy, and a pleasant voice answers, welcoming you to the wonderful world of Acme Drugs, and requests that you ‘please listen carefully to your options’, adding that ‘you may enter the number of your choice at any time’. When the voice finally gets to the option you’re interested in—‘to refill a prescription, press “4” now’—you follow the instruction: you press ‘4’. The voice then says, ‘please enter your prescription number’, and so on and so forth. What is actually going on, of course, is that the computer running the menu is simply using a look-up table like the one described above. The computer takes your typed signals as input, and activates a particular taped message in response. When all goes well, the whole experience is not unlike a conversation with a real person. A naive user might even think that she actually is having a conversation with a real person.

But if you are an experienced user of such technology you know that things often do not go well. Sometimes the menu of options is too limited, and doesn't include one that covers your needs. In that case, the voice may just stupidly—but pleasantly!—keep repeating the inadequate list over and over until you give up. Slightly more user-friendly systems will, after a round or two, have the pleasant voice announce that ‘your call is being forwarded to an operator’ (i.e. a real person). Another problem that can arise is when your inputs don't meet the ‘expectations’ of the program—you enter too many digits, or too few, or you pause too long in between entering. If there is voice-recognition software involved, the machine may find your accent or pronunciation pattern too unlike its stored exemplars to be able to find a match. In such cases, the machine may generate (in a pleasant, pleasant voice!) the message ‘I'm sorry—I didn't understand. Could you re-
Thinking

enter/repeat your information?’. Even if one of these expedients works—and sometimes none does—you are left with the sour knowledge that the call could have been handled in seconds if you had only been allowed to speak with someone who could think.

Returning to our master machine, we can now ask what would happen if the interrogator typed a question that was completely off the wall, or badly mistyped? Block points out that the basic programming strategy outlined above can be both simplified and improved (for purposes of passing the Turing test) if the programmers of our master machine keep a real person in mind as they craft the machine’s responses. Block suggests his Aunt Bubbles: someone who apparently took a slightly dim view of academic exercises, and who thus would be prone to respond to strange or ill-formed questions with a throwaway remark like, ‘My nephew told me you might try to confuse me’. The programmers could take a few remarks such as this, and have the machine cycle through them at random whenever it received an input that did not match any of the anticipated intelligible remarks. This expedient not only allows the programmers to give the machine a plausible and coherent ‘personality’, but provides them with a way of coping with inputs that might otherwise expose the program and expose the deception (lock 1995).

Here, too, there have been working programs that run on such principles. A program written by Joseph Weizenbaum, called ELIZA, was designed to mimic the behaviour of a Rogerian psychotherapist (Weizenbaum 1976). Therapists in this tradition are ‘non-directive’—they do not offer their patients analysis or advice; instead they try to facilitate their patients' self-discovery by asking open-ended questions prompted by themes in the patients' own comments. This fact about Rogerian therapeutic practice makes it ideal for mimicry by Aunt Bubbles-type programs. After a brief opening question, ‘Why have you come to see me?’ or ‘How are you feeling today?’ the machine can let the ‘patient’ take the lead, responding with highly stereotypical responses, sometimes picking up on words in specified grammatical positions. Here’s a typical example:

ELIZA: WHY HAVE YOU COME TO SEE ME?

Patient: I have been feeling depressed.

ELIZA: CAN YOU ELABORATE ON THAT?

Patient: Everything I do is wrong.

ELIZA: CAN YOU BE MORE SPECIFIC?

Patient: Well, my boss is very unhappy with my work.

ELIZA: TELL ME MORE ABOUT YOUR BOSS.
Although literature and lore are full of anecdotes about people who stumbled on to ELIZA unawares (and this is easier to do now in the age of email and the internet than it was back in the 1970s) and were fooled into thinking they were communicating with a real person, the program was pretty easy to unmask. But with Block's Aunt Bubbles machine we have the guarantee of a much more effective deception. As long as the interrogation lasts no longer than an hour, the machine will pass the Turing test. And of course, following Block's principles, we could design a machine that would last even longer by expanding the size of the original list. But while this machine would readily fool a human interrogator, no one who knows how it is accomplishing the deception would say that the machine is thinking. What Block's example shows is that we are not willing to identify thinking with whatever happens to be responsible for apparently intelligent behaviour. We want more than just intelligent-looking behavioural responses—we want the responses to be caused in a certain way.

But this, of course, is just the point that Descartes emphasized in his Discourse argument. Behaviour can appear to be the result of thinking, without there really being thinking behind it. Notice that Block's machine is a version of the kind of physical device Descartes thought was 'morally' impossible. That is, the Aunt Bubbles machine essentially has a special-purpose 'organ'—in this case, a specialized subroutine—available for every envisioned circumstance. What's important to notice is that such a machine is not literally impossible; it is only practically impossible. Merely writing down the program we've been talking about would take more time and material than is available to any human being in the space of a lifetime. The amount of time it would take a physical device to search the relevant lists would also be prohibitively large. What all this means is that the look-up table is not a plausible explanation of the intelligent behaviour we observe in other human beings. While Descartes may have been wrong about the capacities of matter in general, he was absolutely right in thinking that no human head—indeed, no physically possible being, period—could produce the range and amount of intelligent behaviour that we do by relying on a bag of special-purpose tricks.

We have, then, a happy convergence between an intuitively appealing condition on intelligence—that it must involve a systematic capacity—and an empirical argument that systematicity is the most likely explanation for the intelligent behaviour we actually observe. But is systematicity all there is to intelligence? It seems that the answer to this question is clearly no. After all, ordinary computers are systematic in their operation, but few people would say that these devices genuinely think. What's missing?

To see what's missing in our account of thinking, let's look back at the Aunt Bubbles machine. Imagine that in place of a machine that's programmed to match remarks we instead achieve our effect by employing a human being to sit in a room and follow the instructions that constitute the machine's program. So our Aunt Bubbles operator will go through the procedure we imagined the machine doing automatically: she will look up the input strings, match them to conversations, and output the next appropriate strings. This is, of course, a rather silly procedure for a human being to go through, given that she could just read the input remarks and then think of appropriate replies. But now let's change the case a bit. Imagine that the conversations, inputs, and outputs are all in Chinese. If our operator is a monolingual speaker of English then she will not be able to understand any of the remarks she is dealing with, but—and this is the significant thing—
she will still be able to manipulate the strings of Chinese characters in accordance with the instructions, so as to simulate just as effectively as before the behaviour of an intelligent person.

The set-up I’ve just described is—except for one wrinkle that we’ll add in a minute—identical to the set-up devised by John Searle in his landmark essay, ‘Minds, Brains, and Programs’ (1980). His thought-experiment, now known as the ‘Chinese Room’, was devised to show that the limitations of machines go beyond those identified in the Aunt Bubbles example. According to the line of thought we were pursuing, the problem with the machine, the thing that kept it from being a genuine thinker, was the fact that it relied on the wrong kind of program: it used a ‘stupid’ look-up algorithm instead of a program that exploited the compositional structure of the remarks it was manipulating. But according to Searle, the problem is the fact that the machine is running a program, period. A program is a set of instructions for the manipulation of symbols. Crucially, the instructions treat the symbols as mere formal objects, things that can be identified purely by their shapes. As far as the instructions go, the symbols may mean anything, or nothing. Consequently, the instructions can be followed by someone—or something—who has no idea whatsoever what the symbols mean.

Much earlier we noted that thoughts have contents, that they represent or are about things, that they possess the property philosophers call intentionality. But of course thoughts are not the only things that can represent. Ordinary objects can stand for things, as a red light stands for the command ‘stop’, and gestures can too, as a thumbs-up gesture is a symbol of approval. And, of course, words stand for things. But there’s a difference. Objects, gestures, and words, it appears, derive their intentional contents, their meanings, from the thoughts they function to express. Because we are thinkers, we can impose meaning on things—light fixtures, gestures, vocalizations, and squiggles on paper—that are otherwise completely insignificant. Because of conventions adopted by our ancestors, the word ‘horse’ is used by some of us to refer to horses; other people, because of different conventions adopted by different ancestors, use the word ‘cheval’ for the same purpose. The intentionality of words and other public symbols is thus a derived intentionality. The source from which this intentionality derives is the original intentionality of thought.

Now of course we can impose derived intentionality on anything we like, including the symbols processed by computers. And we do. We speak of computer data as ‘code’, in recognition of the fact that we can construe the strings of ones and zeros that we type into the machines as representations of any information we want to store or utilize. And we can write programs—instructions for the manipulation of code—that mirror in form the thought processes that we want to apply to that information. That’s what Turing showed us we could do, remember: devise a mechanical device the operation of which mirrors any thought process that can be carried out in a step-by-step way. But to say that the mirroring is formal is to say that the operations are sensitive only to the formal
properties of the symbols, that they would work the same way no matter what the symbols mean. Indeed, as far as the operation of the machine is concerned, the symbols mean nothing at all, just as the Chinese symbols manipulated by our Aunt Bubbles operator mean nothing at all to her.

Do these considerations show, as Searle contends, that thinking cannot be computation? Defenders of the computational account have many criticisms of the Chinese Room example. To begin with, consider the program the Aunt Bubbles operator is following. Lack of systematicity was enough to rule out the original Aunt Bubbles machine as a genuine thinker; mightn't that defect be the problem in the new case as well? Moreover, we know that no actual speaker of Chinese could be following the program that the original machine follows, because there is not enough space in a human head to store the needed data, nor enough time in a human life to effect the required searches. If speaking Chinese involves, as the computationalist assumes, running a program, that program is very different from the one run by the Aunt Bubbles machine. What happens if we give that very program to our friend in the Chinese Room?

This is the wrinkle that I referred to above. Go ahead, says Searle, give the person in the Chinese Room whatever set of instructions you like, including whatever program it is you think characterizes the competence of a Chinese speaker. The result, he contends, will be the same. The instructions will still always treat the words of Chinese as meaningless squiggles; the operator will always be able to follow them successfully without understanding a thing. Indeed, Searle adds, we can even imagine that the operator memorizes the program, so that her instructions are completely internalized, as is, presumably, the program 'running' in the head of a Chinese speaker. The intuition persists: our operator does not understand Chinese.

Another response made by defenders of the computational model concerns the way in which the operator in the Chinese room interacts with the outside world—or, rather, the way in which she does not interact with the outside world. Actual thinkers can gather information about the external world by means of their various sense organs, and can affect that world by moving their limbs. It’s this causal interaction with the external world that enables a thinker to attach meaning to the symbols it manipulates in thought. But once again Searle is happy to modify the set-up to remedy the claimed deficiencies. Give the operator a camera, through which she can see the inputs being delivered, and let the symbols she outputs be wired up so as to cause movements in artificial limbs. This is analogous to attaching a computer to a robot, says Searle, and makes no difference at all to the crucial intuition. We would have, in this case, a monolingual English speaker causing the movements of a robot in ways that are appropriate (we may assume) to sentences in Chinese, but not via an understanding of the Chinese characters.

But if Searle is right, that nothing we can add to the operator’s situation will make the manipulated symbols come alive with meaning, where does intentionality come from? What makes it the case that thoughts represent things outside themselves? Searle’s answer to this question is notoriously obscure. Intentionality, he says, is a biological
product of the human brain. In saying this, he insists that he is not ruling out the possibility of an artificial brain; that is, one made of non-biological materials. Such a ‘brain’, he says, could even be made out of electronics, provided only that it possesses the same causal powers as a human brain (1980, 1992). But it's difficult to understand what Searle has in mind. The brain has many causal dispositions; an artificial brain would presumably need to possess only some of these. Which causal powers would be the ones essential to the ‘production’ of intentionality, and which ones could the designer of an artificial brain neglect? In the case of artificial hearts, it is the causal powers that have to do with the heart's function that we want preserved. Those that have to do with its reaction to, say, serum cholesterol are not properties that we want preserved; indeed, they’re ones we’ll take care to avoid duplicating. So it's presumably those properties of the brain relevant to the function of thinking that are the important ones. But if we use functional criteria to single out those ‘causal powers’ of the brain that are crucial for intentionality, then it's not clear how Searle's view differs from the functionalist's after all.14

But in any case there is a point about the Chinese Room thought-experiment that we are neglecting, one that's key to understanding how thinking might be computation after all. Consider the fact that while the Chinese Room operator does not understand the Chinese symbols she is manipulating, she does understand something; namely, the English instructions. This suggests that the Chinese Room paradigm may not provide a fair test of the computationalist's hypothesis. Since the operator's relation to the English instructions she understands is obviously functionally different from her relationship to the Chinese symbols she does not understand, it’s open to the computationalist to say that the thought-experiment does not capture the relevant relationships between operator and language—that whatever program the operator is running, it is not the ‘program’ that human beings ‘run’ in the course of understanding language.

While we’re at it, let's look back at the Aunt Bubbles machine. Here again, on a second look, we have some reason to revise our original assessment of the situation. For the machine to carry out the ‘stupid’ operations associated with the look-up table, operations, that is, that require no understanding whatsoever of the English strings it is receiving as input or producing as output, it must be running some program that enables it to respond appropriately—non-stupidly—to the symbols in which the program is written. We have already encountered the concept of a ‘machine language’. This is the code in which the most basic computer commands are written; it's the symbol system that directly triggers changes of states in the computer. Perhaps if we could get clearer on what it is for the machine to figuratively understand its own machine language, we could get an idea what it is, in computational terms, for the Chinese Room operator to really understand her instructions. When we figure that out, perhaps we'll be able to characterize the difference, in computational terms, between the operator's relationship to the English instructions and her relationship to the Chinese characters that she does not understand.
Thinking

Let's pretend that the Aunt Bubbles machine not only stores an enormous list of conversations expressed in English, but that its machine language is a simplified form of English. In that case, we could give the machine the following instruction: ‘If FIND “FIND” THEN SAY “OK”’. There are two occurrences of the English word ‘FIND’ in this command, but they differ importantly in the effects they will produce in the computer. The first ‘FIND’ will cause the machine to initiate a sequence of state changes that constitute—because of the programmer’s cleverness—a search routine. The second ‘FIND’, however, because it is enclosed in quotation marks, will not instigate such a sequence. The computer's program will treat this occurrence of the symbol as a mere object, and not as the expression of a command. That's the effect of the quotation marks: they cordon off the symbol and present it to the computer as a meaningless, inert shape. The fact that the shape is the same as the shape of the symbol that, when unquoted, prompts the computer to begin a search is invisible to the computer, and irrelevant to its operation. The machine's program uses the symbol in one case, and merely mentions it in the second.

Functionally speaking, we can mark this distinction by saying that the unquoted symbol possesses, for the machine, a conceptual role, a distinctive pattern of causal relations to the computer's internal states, to its other potential inputs and outputs, and, in particular, to the other symbols that constitute the machine's machine language. The quoted symbol, on the other hand, has no such distinctive causal profile (although the quotation marks themselves do). This, then, is the difference between the Aunt Bubbles machine’s relation to the symbols that compose its instructions and its relation to the symbols that it is instructed to manipulate: the former, but not the latter, have a conceptual role in the machine's functional architecture. It can use the symbols that compose its instructions, but it cannot use the symbols that are merely mentioned in those instructions. The same distinction can be drawn, arguably, in the case of the Chinese Room. The symbols that compose the operator's instructions (whether written down or vocalized to her) must be integrated into the operator's overall psychology in much the same way that the ‘FIND’ symbol is integrated into our machine's internal functioning in order for her to understand them. The fact that the Chinese symbols are merely mentioned, and thus the fact that they have no conceptual role within the operator's psychology, is sufficient to explain why she does not understand them.

But is conceptual role all there is to meaningfulness? To put the question another way, must we now say that the Aunt Bubbles machine really understands its instructions? We do not. For there is a difference between the kind of conceptual role possessed by symbols in a machine's machine language and the kind possessed by mental representations in the minds of real thinkers. The difference has to do with the way in which the symbols are connected—or not connected—to states of affairs in the external world. In the 'conceptual economy' of a mere computer the internal relations among the internal states of the machine are independent of events occurring in the external world. All of these states are interconnected with the inputs and outputs the machine receives,
but which inputs the machine receives, and the effects of the outputs the machine produces, are unrelated to anything but the intentions and interests of the person using the machine.

In contrast, the inputs that a genuine thinker receives are information-bearing: they co-vary reliably with specific external states of affairs. Similarly, the outputs produced by a genuine thinker—typically, actions of one sort or another—have consequences in the external world. These patterns exist independently of the will or intentions of any thinker; they are matters of natural law. Because the conceptual roles of symbols in the internal representational system of a genuine thinker necessarily involve a coordination of these ‘naturally’ meaningful inputs and these consequential outputs, the ‘interpretation’ of the symbols will be naturally constrained. I can impose whatever interpretation I want on the symbols on which a computer is operating, treating the one and the same configurational state, perhaps, as a position in a chess game on one day and a depiction of erosion in Iceland the next. In so far as such a state has a meaning at all, it is a meaning that derives from my intentions to use it in a certain way. I am not similarly at liberty with respect to the meanings of my mental symbols. DOG will mean ‘dogs’ in my psychological economy just in case the occurrence of that symbol is connected in lawful ways, via its connections with information-bearing input states and consequential output states, to environmental circumstances involving dogs. The intentionality, or meaningfulness, of my mental symbols thus originates in the non-intentional, natural world.

To say that the mental symbols deployed by real thinkers possess original intentionality is not to say that the intentional contents of the symbols affect the operations of the computational device that realizes the thinker’s mind. Fodor has argued that computationalism is committed to a principle he calls methodological solipsism: the principle that computational mechanisms are sensitive only to intrinsic, formal features of the symbols they manipulate, and a fortiori not to the relational properties that determine their meanings. But methodological solipsism, Fodor insists, does not entail that the manipulated symbols have no meanings. He charges that Searle makes exactly this erroneous inference, conflating formal operations on symbols with operations on formal symbols. To say that the operations that a computer (or, as computationalism would have it, a brain) performs are formal is to say that the causally relevant features of the representations over which the operations are defined are formal features. This is just the doctrine of methodological solipsism, and is indeed something to which computationalism is committed. But it is no part of computationalism to say that the symbols on which the mental operations work are formal symbols; that is, symbols devoid of meaning (Fodor 1991).

Here then, finally, is a characterization of thinking: thinking is the operation of structure-sensitive processes defined over the compositional structures of symbols possessing original intentionality. It falls out from this characterization that thinking is something that a purely physical entity can do, but not something that only a biological brain can do.
I will leave matters here. In doing so, I acknowledge that I leave many objections unanswered, and many questions unaddressed. In particular, I have said nothing about the relationship between thinking and consciousness, a lacuna many philosophers—including both Descartes and Searle!—will find unforgivable. Let me just assert—without argument, and in the face of many contemporary defences of his view—that Descartes's fundamental mistakes about the capacities of matter carry over, and call into question his arguments for the connection between thinking and consciousness. I do not mean to cast stones at Descartes, however. My own characterization of thinking was developed on the basis of philosophical argument and low-level empirical observation—the very kind of armchair theorizing that Descartes undertook in his own *Meditations*. Still, I believe that my account comports nicely with our emerging scientific understanding of the mind, with work in cognitive psychology, linguistics, ethology, information sciences, and neuroscience. Just as Descartes would have expected.

**APPENDIX: Empiricism and Connectionism**
Thinking

As I remarked earlier, the conception of thought I've been presuming is the Cartesian, rationalist conception, which treats inference as the paradigm mental process. But it must be acknowledged that much of our mental life does not fit this conception. In particular, a good portion of our thinking involves non-rational trains of thought. For example, I hear news on the radio of the Iowa caucuses. Thinking of Iowa makes me think of my favourite musical, *The Music Man*, which is set in Iowa. The phrase ‘music man’ makes me think of the phrase ‘confidence man’, which reminds me of a magazine article I read about confidence men in the mid-nineteenth century, which makes me think of the bill sitting next to the magazine that I need to pay, which reminds me that paying the bill will trigger an overdraft.

In such cases as this the thoughts composing the train of thought are not bound by logical relations among the thought contents, but rather by loose relations of some kind of similarity. They are bound by association. Against Descartes and other rationalist theorists, the British empiricists—John Locke, George Berkeley, David Hume in the eighteenth century and John Stuart Mill in the nineteenth—held that all thinking, fundamentally, had this character. In their theories they sought to characterize the principles that explained why certain ideas would be associated with others, principles like resemblance (Iowa caucuses and the Iowa setting, ‘music man’ and ‘confidence man’), contiguity (the magazine and the bill), and causation (bill paying and overdrafts).

Not only did the empiricists think that associations bound thoughts together in the mind, they held that association bound thoughts to their contents. According to these philosophers, individual ideas got their contents through association with things in the outside world. My idea of a dog was a dog-idea, an idea that meant ‘dog’, because it was an idea of the type that was regularly occasioned in my mind by perceptions of dogs. (This view about the origin of semantic content represents an unusual point of convergence between empiricists and contemporary rationalists like Fodor.) Because of this idea–world connection, relations like similarity and contiguity among objects in the world would be tracked by the ideas thinkers associated with them. In this way, they thought, we get an account not only of thinking, but of knowledge. Properties that are frequently co-instantiated, and event types that are ‘constantly conjoined’ (Hume's phrase) in our experience, and hence in nature, will be mirrored by strongly associated ideas in our minds.

Associative processes do seem to be part of our mental life. Memory, for example, appears to be organized in accordance with the sorts of principles posited by the empiricists. And it’s clear that the content of our stereotypes—mental templates that we rely on for ‘quick and dirty’ conclusions—are built out of experience in much the way the empiricists held that we constructed general ideas from particular experiences.

But empiricism fails as a general account of thinking in offering no account at all of the kinds of thought processes the rationalists focused on; that is, rational processes. From the thoughts ‘Clinton accepts money from big insurance companies’ and ‘No one who accepts money from big insurance companies will implement real health-care reform’ we
might infer ‘Clinton will not implement real health-care reform’. Trains of thought such as these cannot be explained just by association. The last thought certainly is similar enough to each of the preceding thoughts for it to have been triggered by association with them—that's not the problem. The problem is that there are any number of different thoughts that could be equally strongly associated with the previous two, including the negation of the last thought: ‘Clinton will implement real health-care reform’. The relation of association simply places no constraints on the kinds of thoughts that can be associated with each other. (Since contiguity is one of the relations that gives rise to association, we can easily forge an association between any two thoughts whatsoever, just by juxtaposition.) Logic, however, does impose constraints; not just any thought follows—in the logical sense—from any other.

The inability of empiricist models to account for logical inference is connected with the empiricists’ failure to recognize the existence, or at least the importance, of the compositional structure of thought. Most importantly, they failed to apprehend what Frege noticed—that the component parts of thought differ in character from each other. This meant that they had no materials with which to build conceptual structure other than appending and abstracting, and those are not enough.

Just as the von Neumann machine provides a natural model for the rationalistic conception of thought, a different kind of computer, called a connectionist network, offers a way of realizing the empiricist conception. A connectionist network consists in layers of interconnected nodes. Each node can send signals to any node in the upper layer to which it is connected. Every node has a firing threshold, which means that the node will fire when the total strength of the signal reaching it (got by adding the strengths of all the signals) reaches that threshold. Connection strengths can change over time, partly as a function of the frequency with which a connected node fires. The first layer of nodes is called the ‘input layer’, and the last is called the ‘output layer’. Connection strengths between nodes can change in response to the frequency and patterns of firings from the nodes at the level below. If and when a network reaches the point at which new inputs fail to produce any change in connection strengths, the network is said to ‘settle’.

Connectionist networks are good at integrating and propagating information, and at extracting patterns—at performing, in other words, exactly the kinds of mental operations empiricists took to be constitutive of mental life. The following example, though unrealistically simple, should give an idea how this might work. Imagine a network with an input layer in which each node is selectively sensitive to the instantiation of some sensory quality. The network can gain ‘experience’ through activation of its input nodes by signals sent from the relevant sense receptors. As the network's experience accrues, it will begin to become sensitive to patterns of co-occurring sensory qualities, and will eventually begin to produce a distinctive output whenever such a pattern occurs. These
outputs can then be thought of as concepts, constructed by the association of and abstraction from primitive sensory ideas.

Connectionist networks are proving to be valuable as models of those aspects of thought that conform to empiricist principles: memory, some kinds of perceptual processing, and some kinds of associative learning. Many philosophers and cognitive scientists, especially those who are opposed to the computational/representational model of mind promoted in this chapter, contend that connectionism can offer a comprehensive account of thought (see Rumelhart and McClelland 1986; Smolensky 1988; Ramsey, Stich, and Garon 1991). But some philosophers contend that these networks are inherently unable to model those kinds of thought that depend upon the exploitation of compositional structure (Fodor and Pylyshyn 1988; Fodor and McLaughlin 1990; McLaughlin 1993).

References


Thinking


Thinking


Notes:

(1) See Ch. 3 above for a discussion of substance dualism.

(2) I’ll use capitals when I’m indicating thoughts or concepts.

(3) Which is, of course, not to say that I cannot distinguish Hershey’s-kiss-chocolate from other kinds of chocolate, like for instance Ghirardelli-chocolate. I’m no philistine.

(4) It’s an unfortunate fact, and the source of much confusion, that this term is homophonic with Brentano’s term ‘intentionality’, especially since there’s almost complete overlap between the things that are intentional, and the things that are
intensional. For a comprehensive discussion of the meanings of these terms, and the connections between them, see Crane (2001).

(5) Several of the chapters in this volume address the issue of whether qualitative mental states such as sensations have intentional contents, so I will not consider it here.

(6) Frege may have been the first philosopher to recognize this fact, or at least the first to understand its crucial contribution to the compositionality of thought (1892/1970).

(7) For a detailed history of the development of computation and its implications for the study of the mind see Harnish (2002).

(8) There are different varieties of functionalism; one variety, called ‘analytical functionalism’ makes no appeal to a machine analogy. For discussion see Block (1978).

(9) Turing's one doubt about this came, bizarrely, from his reading of the evidence for extra-sensory perception, which he found, ‘at least for telepathy’ to be ‘overwhelming’ (see Turing 1950: 453–4).

(10) Turing’s actual set-up is more complicated, but in ways that add nothing to the structure of the experiment (see Turing 1950: 433–4).

(11) For a complete description of the machine, with some helpful illustrations of its operation, see Block (1995: esp. 381–4).

(12) To visit ELIZA yourself, just google ‘Eliza computer therapist’ and go to any of the dozens of sites that come up.

(13) Depending on the amount of time we want to allow the interrogation to run, it might become physically impossible—we could quickly get to the point where the number of characters needed to write the program would be larger than the number of electrons in the universe.

(14) For fuller development of this criticism of Searle's account of intentionality see Antony (1997).

(15) Actual machine languages are cumbersome and unperspicuous. Since simple commands written in English can be readily translated into machine languages, though, our pretence does no harm.

(16) I should point out that while many philosophers of mind agree that original intentionality involves ‘natural meaning’ in some way or other, they disagree violently about precisely what way this is. Fodor thinks that a distinctive pattern of covariance between symbol and world is sufficient, and views the symbol’s conceptual role as simply one way that the covariance can be causally sustained (1987). Harman (1982) and Block (1986) contend that it's the conceptual role itself that constitutes a symbol's meaning. Both Millikan and Dretske think a symbol's meaning is determined by its function, but
differ in their accounts of what determines a symbol's function (Millikan 1984; Dretske 1988).

(17) Many philosophers disagree that the considerations Fodor cites really count in favour of methodological solipsism, and contend that relational properties, like semantic properties, can be causally efficacious. For discussion see the papers in Heil and Mele (1993). Other philosophers, like Tyler Burge, disagree that mental states can be individuated without reference to their intentional contents (Burge 1986; for discussion see Egan 1991).

Louise Antony

Louise Antony is Professor of Philosophy, Department of Philosophy, University of Massachusetts at Amherst.