RICHARD SPENCER-SMITH

THE TURING TEST CONCEPTION OF ARTIFICIAL INTELLIGENCE*

Abstract

In 1950 Alan Turing initiated serious philosophical debate on the idea of artificial intelligence, arguing that artificial thought would eventually be created. He rebutted several appealing objections to this idea - e.g. that machines could never be genuinely creative. But at the centre of his argument was a criterion for the detection of artificial thought. He proposed that we replace the intuitive question 'How would we know whether a machine was really thinking? by this: 'Could a computer pass a certain very rigorous behavioural test?' Although it would be pratically impossible to construct a computer which could 'cheat' at this test, Ned Block has shown that it is logically possible for a non-thinking computer to pass it. This example points us towards a richer conception for deciding whether a machine possesses thoughts.

Artificial Cognition

I take artificial intelligence (AI) to be the discipline concerned with the computational modelling and simulation of cognition, of the cognitive abilities of humans and other animals. The prospect, taken for granted in nearly every science fiction film, of the goals of artificial intelligence being attained, makes the question 'Could a machine really think?' one of the perennially popular philosophical questions of our times. In this paper I wish to discuss the first major

* A version of this paper was presented at the Department of Philosophy, Ioannina University, April 14th, 1992. I am grateful for Erasmus grant STV - 91-U K-4026, which made the visit possible. Some material here originally appeared in 'Il test di Turing e la cognizione artificiale', in *Kos - Rivista di scienza e etica*, 78 (1992), pp. 10-17.

philosophical treatment of this question, by the British mathematician Alan Turing (1912-54). As a preliminary, I want to comment on both components in the concept of AI: *Intelligence*, and *artificial*.

If current practicioners of AI had the chance to rename their discipline, they might well call it 'artificial cognition'. *Intelligence* is slightly misleading, in that it is just one aspect of cognition. Psychologists tend to disagree on what exactly intelligence is, but certain key features are generally agreed. Two important components of intelligence are mental adaptability or flexibility - the ability to modify and adapt behaviour when appropriate; and the ability to learn - to learn from mistakes for instance, and not repeat them. These are relatively high level cognitive abilities, and generally they presuppose the existence of other, lower, cognitive capacities. The ability to learn requires a memory; the ability to adapt to new situations requires some mode of perception - some means of perceiving the environment, of recognising and classifying objects in it. AI is concerned with all these levels of cognition, which is why I would say that the word 'intelligence' - with its suggestion of IQ tests, problem solving and symbolic thought - is misrepresentative of the enterprise as a whole.

In ordinary language, 'artificial' is contrasted with 'real'; 'synthetic' with 'genuine'. This is because, typically, the synthesized item is not designed to replicate all of the properties of the real thing but only a selected range - a superficial similarity of appearance and texture, perhaps. Synthetic grass is not real grass, it is a substitute. But it is not an essential feature of artificial products that they only partially reproduce the original. Water which is synthesized afresh by combining hydrogen and oxygen in a laboratory is just as much genuine water as that which occurs naturally. For what is essential to some substance's being water is that it has the chemical composition H_2O[1]. 'Artificial' tells us that something is an artefact, made by human hand - this reveals its genesis, but need not detract from its essential nature. With neither contradiction nor exaggeration, we could talk of artificial real water. Moving closer to the matter at issue, consider Miller's famous 1953 experiment in which amino acids were produced in conditions possibly duplicating those on Earth at the time when life first appeared. Imagine an extension of such

1. I here assume the Putnam-Kripke view of natural kinds. See e. g. Kripke (1972).

an experiment producing entities which satisfied what are widely accepted as the defining conditions of living organisms - homoeostasis and the ability to replicate. Again, we should accept that those creations are genuinely alive; synthesized but real living things.

It follows that any notion of artificial X can be understood in at least two ways: an *Ersatz,* a synthetic substitute which only mimics some of the features of the original; or something indistinguishable from a real X, apart from its hand-made origin. In the former, weak sense, artificial cognition already exists - in the form of expert systems, object recognition systems, and all the other products of AI. AI as a technology is concerned with producing systems which perform tasks which, when performed by us, require intelligence or some cognitive ability. Such systems replicate some features of the biological original: the ability of a medical expert to diagnose some specific range of diseases, the ability of people to recognize the faces they know, etc. The philosophically interesting question is whether artificial cognition could exist in the strong sense: real cognition, created artificially. The point of the preceding paragraph is to show that the idea of synthetic real thought is not a contradiction in terms. The question of whether it is possible is simply: can we establish that the synthesized item satisfies the criteria for genuine thought? Alan Turing made a famous proposal concerning this, and it is this I now wish to examine.

'Computing Machinery and Intelligence'

Alan Turing made three contributions which were of immense significance for the development of AI. The first was logical: in 1936 he formulated the first rigorous mathematical definition of computation. His notion of an abstract computing machine - a Turing machine - enabled him to prove some fundamental theorems about the nature of computability. His and other's work on the logical nature of computability laid the theoretical foundations for computer science. Turing's second contribution was practical: evolving out of his work in developing code-breaking machines during the Second World War, he was a key contributor to the development of the first real electronic computers. His third contribution was philosophical - in 1950 the philosophical journal *Mind* published his very influential and much discussed paper 'Computing Machinery and Intelligence'. Although the word 'intelligence' appears in the title, and may

have misled some into thinking that he was proposing some kind of intelligence test for compnters, Turing's real concern was with the relationship between computation and *thought.*

Turing first asks us to imagine a conversational game in which three people participate: a man A, a woman B, and an interrogator C. The interrogator is screened off from the other two. He only knows them as X and Y. Each of the players has a different objective. For the interrogator, it is to identify which of X and Y is the man, which the woman; for the man, it is to thwart the interrogator in this, to deceive him; and for the woman, to aid the interrogator. The set-up is illustrated in Figure 1.

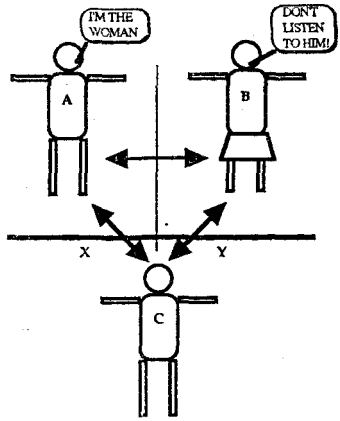

Figure 1. The Imitation Game.
C's aim is to identify X =\A and Y = B, or X = B and Y = A

Obviously, the interrogator should not get any help from physical clues, such as the voices of X and Y, so they communicate only by means of typed messages (interactive electronic mail). Turing called this the Imitation Game; the man wins if he succeeds in *imitating* a woman - or to put a different slant on it, if tricks the interrogator into the false supposition that he is the woman. Dennett says[1] *maybe* the point of the Imitation Game is to suggest that i) men and women think in different ways but ii) we still recognize those different ways of thinking as thinking. I.e. the suggested moral would be that something can think in a very different way to the way you think, but still be recognizably a thinker. That is a nice idea, but it is a rather fanciful interpretation, since this first game was not proposed as a test of being a thinker. Its main point is just to illustrate the

1. Dennett (1990), p. 60.

possibility of a purely intellectual test of imitation, one in which all physical clues have been taken away, so that only the intellectual skills of the participants remain.

The imitation game serves to illustrate the kind of test one could put to a machine. As Turing initially formulates it, he asks: what would happen if a computer takes the part of the man in *this* game - - i.e. will a computer have the same success/failure rate at deceiving the human interrogator C that it is *male* as an average man? But this is misleading. What he really intended - as is evident from his subsequent discussion - is a slightly different game: for the interrogator to identify which of the two correspondents is *human* (Figure 2).

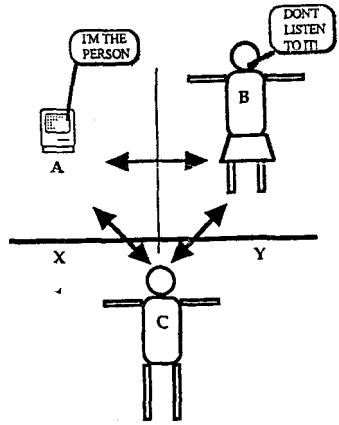

Figure 2. Three-way interaction in the original Turing test.

This is what has come to be known as the 'Turing test':

1) Could a digital computer successfully deceive a human interrogator, under these conditions, that it is human?

This is an *imitation* game - i.e. could a computer successfully imitate a human - or at least, the verbal behaviour of a human under viva voce conditions?

The inspiration for this may have come from the idea of a Universal Turing machine (as suggested later in his article). A Universal Turing machine is his mathematical model of a general-purpose computer - a programmable computer with unlimited memory - so one which, given the right program, can be made to execute any set of computations whatsoever. That is, anything which can be computed at all can be computed by a Universal Turing machine, if it is supplied with the relevant program, the relevant set of instructi-

ons. So if B is a certain kind of machine - a discrete state machine - and A is a Universal Turing machine - then A can indeed perfectly mimic B. With the right software, their behaviour will be *identical* (barring speed considerations). What I am suggesting is that the inspiration for the idea of a game of intellectual imitation may lie in Turing's original logical work on computation. If the human brain were a discrete state machine, then an affirmative answer to question (1) would be inevitable[1]. All you have to do is work out the program (and that is one way of construing what the project of AI is all about). But Turing believed the brain is not a discrete state machine[2]. This does not mean that the answer to question (1) is negative - i.e. that the behavioural repertoire of a human cannot be replicated by a digital computer. It would simply shut one short route to an affirmative answer.

Turing predicted that after 50 years, that is, at the end of the century, there would be a machine with an at least 30% success rate in such a 5 minute viva - the interrogator would detect it 70% of the time. This is better than it might initially seem, since if the machine was so good that it imitated a human perfectly, the interrogator would have to make guesses more or less at random, and thus have on average a 50% success rate. This is an empirical prediction; the timescale has struck most commentators as over-optimistic. But it was surely only a matter of time before someone, in the spirit of those who challenge chess-playing computers, set the Turing test as a genuine competition with a financial reward. Recently the American Computer Museum, in Boston, announced just such a competition. Notice that the test is inherently statistical in at least two respects: not only is a machine subjected to a number of trials, it would also be natural to employ a range of different kinds of judges - a psychologist, a philosopher, an artificial intelligence expert, and e.g. someone who is expert in cross-examination (a barrister)[3].

---

1. assuming physicalism, that there is nothing non-physical about the mind.

2. In the section of the paper devoted to the 'Argument from Continuity in the Nervous System'. As a physical device, the brain undergoes continuous transitions between its states. But the question is whether such a device 'can profitably be *thought of* as being discrete state' (as he himself had put it earlier, § 5). On this, see Block, pp. 23-24.

3. The Boston competition Involved a very restricted version of the Turing test, in which interactions were short and kept to just one topic. Assessment was a matter of degree in this further respect: judges were only required to make a

The philosophically important point was not the prediction. It was that Turing proposed that the question (1) *replace* the intuitive question

2) Can machines think?

He says of this that it is 'too meaningless to deserve discussion'. He offers (1) as a meaningful replacement for it. In part this was probably a reflection of the verficationist spirit of the time, in which questions apparently incapable of empirical resolution were dismissed as devoid of content. (The fact that one might well feel like saying 'Yes the robot appears to be intelligent, answers questions as if it is conscious, but is it *really* thinking, is it *really* conscious?'- this was just what Turing wanted to dispel; it seemed to him there was no sense in pressing that further question. It seemed there is no better evidence that could be brought to bear. Whereas (1) is a question with clear empirical content.) But even without a verificationist attitude to meaning, two words in the intutive question (2) remain troublesome: 'machine' and 'think'. Obviously the question 'Can machines think?' depends upon what counts as a machine, and what counts as thinking. Turing makes a correct point about a simplistic 'ordinary language' solution to this problem. An ordinary-language philosopher might say 'To find out what counts as a machine, we have to examine the ordinary meaning of the word 'machine''. But this just leads to a trivialisation of the philosophical problem - just go out and conduct a statistical survey of how ordinary speakers use the terms. That could never be a substitute for a careful philosophical analysis of the concepts.

So let us consider the question: what exactly is a machine? We can all point to examples: a car, or some piece of equipment in a factory - but what else is included? The word 'machine' suggests a *mechanical* object - an object, the behaviour of which can be explained in terms of the principles of mechanics. In one sense, this conception is too liberal, since all physical objects have *some* behaviour in accordance with the laws of mechanics (as when dropped from a height and allowed to fall to earth). Perhaps this difficulty can be avoided by reference to the object's *primary* functioning. But the resulting definition would still be unsatisfactory. It would now be too restrictive,

---

comparative evaluation of a number of subjects, ranking them in order of humanness, rather than deciding absoiutely, on each trial, whether the subject was human or machine.

since it would preclude an ordinary electronic computer (or a future optical computer) from counting as a machine. It would be more natural to say that a machine is something whose primary functioning is to be explained in terms of the principles of *physics,* since this allows in the other branches of physics, most notably electrodynamics (and optics). But this proposal still leaves too much undecided, since so much turns on the key notion of what can be explained in terms of physics. If we humans can *ultimately* be explained in physical terms then (2) would receive an affirmative answer by reference to us. But this is not what is intended by that question. (Nor would it do to replace the word 'machine' by 'artefact'. If genetic engineering progresses, it might be possible at some future date to artificially re-create nervous tissue. Biological artefacts capable of thought are not what is at issue). Rather than continue to tinker with the definition of 'machine', it is simpler to substitute a precise term: digital computer. This seems to me to be clearly advantageous; we replace a vague term by a precise one which respects the intention of the question: could a device which is a purely computational device be capable of thought?

Much more controversial is the other substitution Turing proposed. Is the test adequate as a criterion of thought? One may quibble about some of the details of the test. (It has to be said that Turing did not approach the paper as a wholly rigorous academic exercise. This was confirmed by Robin Gandy, a former pupil of Turing's, at a conference celebrating the fortieth anniversary of the paper, who said that Turing intended it to be polemical, to stimulate people into thinking about the repercussions of the then new technology of electronic computers). One kind of doubt Turing rightly dismissed - namely, that a computer would give itself away by instantaneously responding with the correct solution to e.g. an arithmetical problem. It is relatively easy to program a computer so that a) it will pause an appropriate length of time and b) it will occasionally generate the kind of mistakes in calculation we are prone to. There are problems with the essentially adversarial (competitive) nature of the test - the fact that it's a three-way conversation, with X and Y competing against each other. If one partnered the machine with a particularly dim witted person, it might be too easy for the machine to outshine him. If on the other hand one partners the machine with a range of people of differing intellects, does one expect the machine to attempt to be more convincingly human than them

all? We can eliminate the adversarial nature of the test by making it a succession of two-way interactions between an interrogator and a subject (human or computer), rather that the three-way interaction of the Imitation Game (Figure 3).

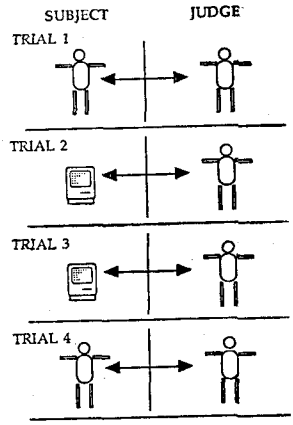

Figure 3. Revised Turing test: a succession of 2-way interactions.

Turing did illustrate the test by means of such a two-way interaction, so he did envisage it in such a modified form. That illustrative dialogue was meant to show that the interrogator could ask anything of the subject, from playing a game of chess to composing a poem. I would expect a successful performance of the test to be not so much an interrogation as a *conversation,* in which the subject throws questions back to the interrogator, challenges his opinions, engages in debate, and so on.

It is important to distinguish two questions about the test:

i) Is passing the test necessary for being a thinker? i.e. *must* anything which can be truly described as entertaining thoughts be capable of passing some form of the test?

ii) Is passing the test sufficient for being a thinker? i.e. does passing the test *guarantee* that the subject can be ascribed thoughts?

One way to think about the first of these is to ask: what would we expect of a person? Suppose someone persistently performed badly at the test. We would not take that as demonstrating that they were devoid of thought. Moreover, subjects cannot pass the test if they do not *want* to pass it. At a different extreme, one might also imagine su-

per-intelligent Martians failing the test because, although condescending to speak our language, they refuse to lower themselves to our intellectual standards, and so are easily discernible through performing *too well*[1]. What such examples illustrate is that passing the test cannot be a precondition for being classified as a thinker. It is not a test which anything possessing thought *must* be capable of passing - thus failure marking one out as lacking in thought. It would leave us with a very anthropocentric conception: Man is the Measure of All Thought. Furthermore, that would make linguistic competence (more precisely, competence with some human natural language) a necessary condition for thought. Although the matter is controversial, many philosophers accept that there can be thought without language. There is nothing to suggest that Turing intended the test to be a necessary condition for thought. All he wanted to propose was a criterion *sufficient* for the attribution of thought: if something passes, it should be classified as a thinker.

We can ask what is the status of the proposed replacement of (2) by (1). It should be clear it is not a descriptive definition - that is, (1) is not intended as a definition of what (2) actually means. As mentioned, (2) suffers from various defects (vagueness, irresolvability) which (1) seeks to rectify. Perhaps then it is a normative definition; (1) doesn't describe what (2) does mean - it shows what it *should* mean, what it ought to mean. As such, Turing's proposal has been said to be an operationalist move. Operationalism (or operationism) is a doctrine in the philosophy of science originally proposed in the 1920s by the physicist P.W. Bridgman, which proposed defining a scientific concept in terms of an operation for detecting / measuring it. It is a rather extreme form of verificationist theory of meaning, i.e. one which equates meaning with empirical conditions of verification - in this case, an operation for empirically detecting something. Such plausibility as operationalism possesses probably derives from the success of the Special Theory of Relativity, which had at its centre just such an operationalist definition: the definition of simultaneity in terms of the operation of sending a light signal.

We can ask whether it is correct in general to define a scientific concept in terms of an operation with a measuring instrument. One

1. Indeed, it is reported that in the first Computer Museum competition two of the ten judges thought one of the human participants was a machine, on the grounds that, for their chosen subject, Shakespearian plays, no-one would know so much about Shakespeare. (*The Guardian,* 28th November, 1991).

problem with this is that it makes the operational test an infallible guide to the presence of the property. The point is not that any single piece of equipment is fallible, can malfunction; we can always calibrate one instrument against others. The point is that by making passing the test definitional, it makes no sense to ask whether e.g. something could have the property in question despite failing the test. And that, as we have seen, is possible in the case of the Turing test. However, there is a way round this objection. In the original form of operationalism, in the interests of prohiblting ambiguity, the the meaning of a scientific term was equated with a *single* measuring operation. On a less extreme form, the meaning of a term could be given by two or more different operations[1]. Once you allow multiple operations, you in effect allow that no single operation is infallible - because one measuring operation might conflict with another. On this view, a single operational test is not definitional but *criterial* - it provides a defeasible criterion for the presence of the property. This is how Dennett views the Turing test. He argues that one could devise other tests for the presence of thought - e.g. conld the subject work out a means of stealing the British crown jewels without the use of violence - but that it is doubtful if one could improve upon Turing's test as a quick, fair, repeatable and reliable test for the presence of thought.

Turing's discussion of Objections

Having set out his main proposal, that question (2) should be answered in terms of the test, question (1), Turing goes on to answer a number of potential objections. But these are not all objections to the same thing. Some are objections to the main proposal, but some are objections to the idea that machines can think (the issue he had earlier dismissed as too meaningless!) Some objections are to the idea of attributing to machines other human qualities (other than thinking, that is). One he calls 'Lady Lovelace's Objection', in honour of the person who originally voiced it, Ada Lovelace (friend of the nineteenth century mathematician and inventor Charles Babbage, who invented a truly *mechanical* computing device). The disputed quality in this case is originality; her objection being that a

1. On this contrast, compare Hempel, p. 126.

computer cannot possess it, it can only do 'whatever we know how to order it to perform'.

It is possible to distinguish two quite different versions of this objection, two quite different rationales which might prompt someone to make it. One version has an epistemological slant; it is concerned with how we perceive the machine, or how it appears to us, irrespective of what is going on inside it. The other has a more ontological slant, concerning how the machine is in itself. The epistemological version is the weaker of the two. Its central point is that a machine could never really surprise us. This kind of argument is likely to appeal most to those with little experience of computers, and Turing makes the correct reply, that machines take us by surprise all the time. Anyone who has done even a small amount of programming is likely to have been surprised by the difference between the expected and actual performance of their programs.

Someone who accepts this version of Lady Lovelace's objection may dispute this. They may argue 'The only way a programmer could be surprised by a machine they have programmed would either be through a failure in their understanding, or a lack of effort on their part, or a malfunction in the machine. It may be that the programmer does not understand the program properly; this is especially common with those who are still learning to program. Or it may be that the programmer does not make the effort to do the calculations; if they did take the trouble they would know its consequences exactly. Or it may be that some hardware malfunction causes the machine to misbehave. But none of these are interesting examples of surprise', the objector concludes, '*in principle,* a machine's behaviour is totally predictable from the program'[1]. As Turing points out, this kind of view is probably founded on a mistaken conception of belief, that you believe all the consequences of your beliefs. The idea is that belief is closed under the relation of logical consequence: if you believe that p, you also believe everything that follows from p. Thus the objector in effect says: if you know some program, you also know all its consequences - running the program will not tell you anything new. But this view is demonstrably false. There are many sorts of problem (puzzles, arithmetical and logical pro-

1. I ignore the possibility of introducing a randomizing element into the program. If randomness or pseudo-randonmness is introduced into the machine, the appropriate notion of prediction would be statistical; taking that qualification into account, the machine's behaviour would still be predictable.

blems) for which the answer is a logical consequence of the problem description, and for which there are people who fully understand the problem but who *cannot* work out the answer[1].

The other version of Lady Lovelace's objection is more concerned with who bears responsibility, or deserves the credit, for the machine's performance. The thought here is that whatever behaviour the machine exhibits, that is only a result of it executing whatever program we have given it. This version of the objection does not essentially concern whether the machine takes us by surprise or not; rather it concerns the source of any creativity it exhibits. If a machine exhibits some psychological characteristic C (whether that is creativity, success at the Turing test, the ability to play chess at grandmaster level, or whatever), and that machine's behaviour is due to its having been programmed by us, then the real credit for C lies with us, not with it. (In just the same way as if person *a* formulates a detailed set of instructions which, by following them, person *b* draws a pretty picture, the credit for devising the picture lies with *a*. *b* is no more than an elaborate tool, a means for *a* to produce a copy of the picture). The correct response to this, I believe, is to concede this point, but to deny that this means that computers cannot be original. For not all computer behaviour is the result of explicit programming by us.

This second version of the objection ignores the possibility of machine learning. There has been a lot of research on machine learning within both of the major traditions of AI, the symbolic and the connectionist paradigms. On the symbolic approach, the problem is to design procedures which will enable a machine to take a set of data and formulate useful concepts about those data, extrapolate projections, formulate conjectures[2]. Within the connectionist tradition, the training of artificial neural networks is perhaps the major research topic[3]. The knowledge of a neural network is not contained within a set of explicit symbolic instructions, but in the pattern of

---

1. What may be true is that if you sincerely believe a proposition p, you are *cemmitted* to all the consequences of p. What this perhaps means is that if you were perfectly rational, and had enough time and memory, and someone *gave* you a correct proof that some consequence q follows from p, you would accept q. But that is far from saying that you already believe such consequences. We can be surprised by what our commitments are.

2. See e.g. Thagard (1988), Chapter 4.

3. See e.g. Rumelhart and McClelland (1986).

connections between its artificial neurons. To train a network is thus to get it to settle on an appropriate pattern of connections. This raises the prospect, at least in principle if not now in practice, of an artificial neural net being trained to an extent comparable to a human infant's training. Thus, if a machine could properly learn from experience, the responsibility for the beliefs and mental abilities it comes to acquire will not be attribntable to some external human agent (a programmer). No-one would be able to claim 'I gave it those beliefs'. If such a machine comes to attain intelligence and creativity, this will be due to its general educational history, to its interaction with its teachers and its environment.

Another objection Turing considered, which gets more to the heart of the matter, is what he called the Argument from Consciousness. He quotes from a certain Professor Jefferson, who argues 'Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt... could we agree that machine equals brain'. Turing counters that, if taken seriously, this position leads to solipsism: the only way to tell for sure if a machine - or a person - is thinking or feeling emotion is actually to *be* that machine or person. That assurance is something each of us only has in our own case. You will know that, if indeed it is true, and you can tell others that you are conscious - but such verbal pronouncements are precisely the sort of evidence which this argument rejects as inadequate. The difficulty with this Argument from Consciousness is that it can appear incapable of resolution; there is a danger that its proponents and opponents chase each other round in a circle, with neither side decisively gaining over the other. The objection makes the point, correct from the standpoint of how things are (ontology), that mental states are not constituted by outward behaviour; a pain, for instance, is an inner sensation, it is not the same thing as the behaviour which typically reveals that a pain is being felt. And the response to this, correct from point of view of what we can actually know (epistemology), is that we have no direct access to those inner states (except each in our own case): the only evidence available is the outward behaviour.

Turing was too hasty in trying to pin a commitment to solipsism on the Argument from Consciousness. Solipsism makes most sense on a dualist view of the relationship between mind and body. On a view which rejects the dualist separation of mind and body - some form of *physicalism* - a person's mental states are only possi-

ble because of the complex physical organ which is their brain. For a physicalist, the solipsistic supposition that other human beings, with functioning brains of similar complexity, might not be conscious, is the puzzling one, because it postulates a difference (in respect of mind) where there is no relevant difference (in the physical basis). But while this may help in the case of human beings, it offers no immediate help with machines. Since the physical composition of a present day digital computer is different to a brain, there can be no argument based on similarity of physical structure. (That is, you cannot argue: consciousness in us requires the presence of a certain physical structure, here's a similar physical structure in this machine, therefore the hypothesis that it is conscious is more plausible than the hypothesis that it is not). So, after all, it might appear that we are back to behavioural output as the only possible source of evidence in the case of machines.

Turing's conception assessed

The Turing test is behaviourist (although Turing does not use the term). Consider again the Imitation Game. In that, successful imitation is successful *deception*; we maintain a distinction between reality (the man...) and appearance (...presenting himself as a woman). But, Turing wanted to insist, there is a difference between the physical and the mental. Unlike the physical characteristic of sexual identity at issue in the Imitation Game, in the Turing test there would be no difference between behavioural appearance and mental reality. The test focuses on a narrow range of behaviour: linguistic behaviour, and by no means all of linguistic behaviour. (Although peson-to-person dialogue is perhaps the most central use of language, there is also, for instance, the use of language to guide action, through commands, requests, and so on). The point, for Turing, was that language could be used as a probe, a window on the (possible) mental life of a subject. Instead of attempting directly to test e.g. the perceptual abilities of subjects, one can ask them to *describe* what they perceive. And there can be questions about the emotional and inner life of the subject, questions which require common sense and everyday knowledge, questions which probe the subject's aesthetic responses, or imagination. Dennett talks of the quick-probe assumption: 'Nothing could possibly pass the Turing test... without being able to perform indefinitely many other elearly intelligent actions'. And of course there was the additional advantage for Turing that li-

nguistic communication enables the subject to be screened off from the investigator, eliminating the possibility of their physical appearance making it obvious whether they are human or machine.

As Professor Jefferson went on to point out, getting a machine merely to signal an emotion is the easiest thing. That is correct, if it means getting the machine simply to print out 'I'm bored' or 'I'm angry', etc. One can see that the Turing test was devised to overcome this kind of point; by rigourously probing a subject on its emotional responses we would soon be able to detect something which was 'merely signalling'. But between trivial signalling and felt emotions there are other possibilities. Consider the following. A machine is programmed with an encyclopedic knowledge of the emotional and aesthetic history of humankind. Also included in its database is a fictional life story for it. It is provided with a set of inference procedures designed to engage with this database, to infer answers about how it felt about certain events in its past or how it would respond under different imagined circimstances, based on making comparisons of its own life history and the records it has of the responses of others. The machine is designed to survive the most rigorous verbal cross-examinaton, but all that is really happening is a complex consultation of uninterpreted sentences which encode information about the emotions of others. The point is not that the machine is lying; as we have seen, deception is inherent in the setup from the outset. It is that although this machine would be behaviourally adequate to pass the test, the relevant area of its inner life would be wholly absent.

If for emotions, why not for thoughts? Might not a machine pass the test by means of the wrong kind of program? There are examples which appear to advance a decisive objection against Turing's behaviourism here. One is due to Ned Block. Suppose you could have a list of all conversational two-person dialogues in English, of a certain limited size. Every utterance - every contribution to a dialogue - must be less than a certain upper limit (500 words, say, or what can be typed by a person in ten minutes or less. You have to place *some* limit on the length of any given contribution, else the number of dialogues would be infinite. Think e.g. of the infinite number of arithmetical questions that could be asked). The total length of a possible dialogue is also fixed, as the maximum length you want any individual Turing test to run. For instance, you restrict the number of utterances in each dialogue to at most one thousand excha-

nges). Then what you will have is a vast but nevertheless finite totality of dialogues. Suppose further that these can be stored in a huge but also very efficient memory store, arranged alphabetically so that they can be indexed for retrieval as quickly as possible. The machine is going to access these in the Turing test. If it was merely a record of all actual —past and present— dialogues, it would be incomplete for this purpose. If it was a list of all syntactically possible dialogues, it would in general produce nonsense. So you have to imagine something intermediate between these: a list of all *plausible* dialogues, of that size. When the interrogator puts the first sentence to this machine in the Turing test, it accesses the memory for those dialogues which start with that sentence. For a given sentence, there will be a vast number of plausible conversations which continue it. (Imagine the interrogator types in the string 'Who is your favourite film star?'. One can think of a large number of possible replies to that, and then an even greater number of replies by the interrogator - some continuing on the theme of films, some dropping that subject and going on to something totally different). The machine picks at random one of these pre-stored dialogues —that is, one of the dialogues which begins with the question 'Who is your favourite film star?'— and prints the reply it contains. The interrogator then types in something else, and the machine then searches for those dialogues which match the exchange as it now exists, question-reply-question. There will still be an astronomically vast number of conversations which continue that initial sequence, so again the machine chooses one at random and returns the next sentence in it. And so it goes on.

In a sense what Block is envisaging here is the idea employed by some actual programs, like Eliza and Parry, which do not generate their own sentences but select them from pre-stored examples, and generalising that idea to a vast imaginary program. The important point is that the machine is not really figuring out the responses itself — sombody else worked out all the plausible dialogues for it. All the machine is doing is unintelligent retrieval; the intelligent work has been done by someone else. Now, this device is a practical impossibility. There are at least three reasons why it could not be built. One is its production: even if every member of the human race had devoted their life to it, the project of writing out the table would hardly have started. Although finite, the number of plausible dialogues is *very very* large. It is diffioult to estimate exactly, but there may be more conversations than there are seconds in the entire

history of the universe. (Yet the table of conversations would have to be updated daily, in order to engage with such questions as 'What did you watch on television last night?', or 'Which team do you think will win tomorrow's big football match?' For any one of those questions, the machine could plausibly fend it off by saying 'I don't watch TV' or 'I'm not interested in football'. But if it kept on fending off such up-to-date questions, it would become very suspicious). A second reason is storage: because there are so many conversations, there may not be enough physical material in the universe on which to store it - you would use up all the atoms trying to write it all out. A third is access: even employing the fastest parallel implementation physically possible, the time taken to look up the answer would prohiblt the thing from engaging in real conversation in real time. Dennett objects that Block's example is 'utterly impossible'. It is important to be clear about thls. It is not logically or conceptually impossible - but given the physical universe we inhabit it is technically and physically impossible.

As Dennett says, the machine would be defeated by the combinatorial explosion. This is the point that the kinds of tasks that AI tries to produce computational solutions to, the kinds of tasks which the mind solves - whether those are theorem proving, playing chess, recognizing faces - generate a vast number of possibilities (Figure 4).

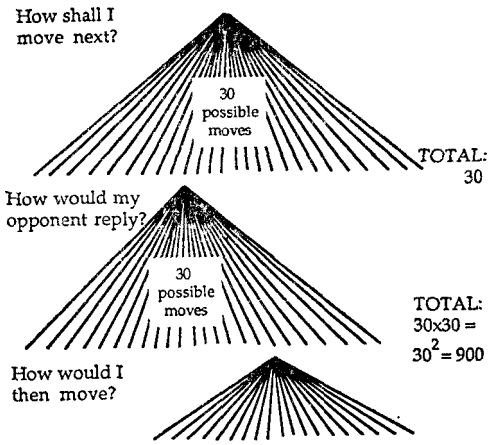

Figure 4. The combinatorial explosion. Suppose that for a given state of the chess board there are, on average, 30 legal next states. If one were to play chess by searching through all possible states, to plan 5 moves ahead would involve 30^{10} (5 moves plus 5 replies) = 590,490,000,000,000 states. Similarly, in looking 10 moves ahead the number of states would be 30^{20}=348,678,440,100,000,000,000,000,000,000.

You cannot solve a task involving combInatorial explosion by laboriously searching through all the possibilities - there are just too many. Brute search through the space of all possibilities is precisely the 'unintelligent' way to go about trying to crack the problem. This is such a fundamental point that it led Newell and Simon to say that avoiding the threat of combinatorial explosion is the task of intelligence. So, Dennett's objection is that Block's program wouldn't work. It just remains a very abstract, logical possibllity - logically, let us suppose, it would pass the Turing test, but it is technically impossible because of the combInatorial explosion. Responding sensibly in conversation is another problem which involves an exponential explosion of possibilities, and which thus requires genuine intelligence - Block's program is not a realistic example.

There is an interesting issue in philosophical methodology about the status of such examples. There is a class of philosophical examples - thought experiments - which are way beyond the realms of science fiction. While logically possible, they are very remote logical possibilities. Some philosophers would say we don't have to pay any attention to them - they are so remote they belong to the realm of fantasy. I would say the point is rather as follows. The rôle these thought experiments are intended to play is similar to that of so-called 'crucial experiments' in empirical science - to provide an experimental test that can decide between two competing theories. In Block's case the rival theories are essentially behaviourism and psychologism. Psychologism he proposes in this weak form - as the thesis that 'whether behavior is intelligent behavior depends on the character of the internal information processing that produces it'. I think that whether one is sympathetic to Block's example depends on which of these two rival theories you favour. Firstly, do you think that the only evidence that is relevant to the mind is ultimately behavioural? And secondly, the more specific issue here: do you think that the only evidence which is relevant to assessing the mental capacities of *machines* is behavioural? Block's example is only intended to demonstrate a conceptual point - that we can judge a machine not to be engaging in thinking even though it is behaviourally adequate.

My own view sides with Block and against Turing and Dennett, for the following quite fundamental reason. The major contribution which computers make to the study of the mind is that they offer precise, computational models of the internal information-processing

involved in different mental functions. In relation to the mental, there is not only the behavioural level, and of course the neurophysiological level - there is also the computational level: what kinds of computations are being carried out in such-and-such mental process? The computer is like a new instrument which enables us to probe inside the mind. Many cognitive scientists would say that in this it makes a fundamental difference to psychology[1]. The mental does not remain an inaccessible inner realm; if you can articulate a psychological theory as a set of computations it can be evaluated very rigourously. So in this I think the Turing test has ultimately been something of a red herring in philosophical discussions of AI: in suggesting that the way to assess the question 'Can machines think?' must be behavioural, it has obscured the key contribution which computers bring to the study of the mental. And, as Block emphasizes, the point is not that in order to think a machine must carry out exactly the same kinds of computations as *our* brains compute. That would be an unwarranted chauvinism. The point of his example is relatively modest: that there are some forms of information processing which we can clearly recognize do not count as thinking. It is not simply that the answers would have had to have been worked out for it by someone else, thus laying it open to the latter version of Lady Lovelace's objection; it is that it does things the wrong way. We can see that the inner mechanism which produces the verbal behaviour is wrong. Operating a vast lookup table is not thinking.

The Block example shows that it is wrong to focus solely on outward verbal behaviour. There are various different possible sources of evidence as to whether some putative thinker does indeed think. There is i) its intellectual behaviour, expressed through language - the kind of thing actually examined in Turing's test; but there are also ii) the inner workings of the subject, its hardware or physiology and software; and iii) its broader outward interaction with the environment. If we ask which of these is closest to the essence of thought, it is probably i). We don't rule out the possibllity of creatures with radically different physiology from ours being capable of thought; nor would we necessarily disqualify a severely disabled person with minimal perceptual and motor engagement with the outside world. But it is also true that i) is not wholly indispensable either. Those who would credit higher animals with thoughts do so, typical-

1. See e.g. Johnson-Laird, Chapter 1.

ly, because there is sufficient physiological complexity required to sustain thought combined with evidence in non-verbal behaviour of primitive ratiocination. Thus it would be wrong to commend the statistical nature of the test simply because the ascription of thought must, in the end, be a matter of degree. If we are to commend the test on those grounds it would have to be that the gradations in passing it correspond to the degrees in which mentality is ascribed. But this is not so. The behavioural evidence on which animals are credited with thought to a lesser degree than ourselves involves, for instance, evidence for the two aspects of intelligence noted above: an ability to learn, and memorize various features of the environment; and an ability to respond flexibly, rather than in ways simply preprogrammed by evolution - as revealed in solving various kinds of problems and, in the case of higher primates, to use tools to manipulate their environment. But all non-human animals fail the test *outright,* for none can master a language.

The point, then, is that thought is not a single, indivisible concept, something which is either definitely present or definitely absent in a thing. There are at least these three aspects to thought and its ascription, and in different cases they are present in differing degrees. It may be that we credit a subject with thoughts when one of them is absent, if there is plausible evidence for the other two, and also no over-riding evidence against. As we have seen, Turing discounted the third of these features, not on the grounds that it is not of the essence of thought, but because he thought it would give the game away. Of course, he was writing at a time when the computer power which today can be slipped inside a briefcase would have required a whole building to house. Today it seems less important that, in judging a computer, we should protect it from prejudice by screening off its physical appearance as that we should be allowed to observe it interacting with its environment[1]. Similarly, Turing did not include the second feature of thought; not because he dismissed it but presumably because he overlooked it. Despite the importance of much of Turing's discussion, the test itself has been unfortunate in that has made subsequent discussion focus too much on the first feature, on the question of the adequacy of a machine's verbal beha-

1. See also Davidson, pp. 6-9, who argues that seeing the thing intereract with the world is essential to discovering its semantics - the connections between its words and their semantic values (things, events, etc.)

viour. The great advantage of the computer as a tool of psychological research is that it enables us to implement theories of the inner workings of the mental, to implement and test models of the internal processes of the mind. With the resurgence of interest in neural networks, the possibility arises of machines whose internal structure and operations are much like those of our brains. In the future, as machines progressively come to compare with us in all these three areas, the conclusion that computers can genuinely think will become increasingly hard to resist.

Bibliography

Ned Block: 'Psychologism and Behaviourism', *The Philosophical Review,* **XC** 1981, pp. 5-43.

Donald Davidson: 'Turing's Test', in K.A. Mohyeldin Said et. al. (eds.) *Modelling the Mind* pp. 1-11 (Oxford: Clarendon Press 1990).

Daniel C. Dennett: 'Can Machines Think?', in Haymond Kurzweil *The Age of Intelligent Machines,* pp. 48-61 (Cambridge: MIT Press, 1990).

Carl Hempel: 'A Logical Appraisal of Operationism', in his *Aspects of Scientific Explanation* pp. 123-133 (New York: Free Press 1965).

Philip Johnson-Laird: *Mental Models* (Cambridge: University Press 1983).

Saul Kripke: 'Naming and Necessity', in D. Davidson and G. Harman (eds.) *Semantics of Natural Language,* (Dordrecht: Reidel, 1972).

D. Rumelhart, J. McClelland, and the PDP Research Group: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* Volumes 1 and 2 (Cambridge: MIT Press 1986).

Paul Thagard: *Computational Philosophy of Science* (Cambridge: MIT Press, 1988).

Alan Turing: 'Computing Machinery and Intelligence', *Mind* **LIX** 1950, pp. 433-60; reprinted in e.g. Margaret Boden (ed.) *The Philosophy of Artificial Intelligence* (Oxford: University Press, 1990).