Artificial Intelligence 60 (1993) 1–22
Elsevier

ARTINT 933

Turing's Test and conscious thought

Donald Michie

The Turing Institute, George House, 36 North Hanover Street, Glasgow G1 2AD, Scotland UK

Received July 1990
Revised May 1992

Abstract

Michie, D., Turing's Test and conscious thought, Artificial Intelligence 60 (1993) 1–22.

Over forty years ago A.M. Turing proposed a test for intelligence in machines. Based as it is solely on an examinee's verbal responses, the Test misses some important components of human thinking. To bring these manifestations within its scope, the Turing Test would require substantial extension. Advances in the application of AI methods in the design of improved human–computer interfaces are now focussing attention on machine models of thought and knowledge from the altered standpoint of practical utility.

Introduction

Although its text is now available in a collection of papers published by MIT Press (edited by Carpenter and Doran [4]), there is little awareness of a remarkable lecture delivered to the London Mathematical Society on February 20, 1947. The lecturer was Alan Turing. His topic was the nature of programmable digital computers, taking as his exemplar the "Automatic Computing Engine" (ACE) then under construction at the National Physical Laboratory. At that time no stored-program machine was yet operational anywhere in the world. So each one of his deeply considered points blazes a trail—logical equivalence to the Universal Turing Machine of hardware constructions such as the ACE, the uses of variable-length precision, the need for large paged memories, the nature of "while" loops, the idea of the subroutine, the possibility of remote access, the automation of I/O, the concept of the

Correspondence to: D. Michie, The Turing Institute, George House, 36 North Hanover Street, Glasgow G1 2AD, Scotland, UK.

operating system. Turing then considers the eventual possibility of automating the craft of programming, itself scarcely yet invented. He further discusses the forms of predictable resistance to such automation among those whom today we call DP (data processing) staff:

> They may be unwilling to let their jobs be stolen from them in this way. In that case they will surround the whole of their work with mystery and make excuses, couched in well chosen gibberish, whenever any dangerous suggestions are made.

We then read

> This topic [of automatic programming] naturally leads to the question as to how far it is possible in principle for a computing machine to simulate human activities . . .

and the lecturer launches into the theme which we know today as artificial intelligence (AI).

Turing put forward three positions:

Position 1. *Programming could be done in symbolic logic and would then require the construction of appropriate interpreters.*

Position 2. *Machine learning is needed so that computers can discover new knowledge inductively from experience as well as deductively.*

Position 3. *Humanised interfaces are required to enable machines to adapt to people, so as to acquire knowledge tutorially.*

I reproduce relevant excerpts below, picking out particular phrases in bold type.

1. Turing on logic programming

> I expect that digital computing machines will eventually stimulate a considerable interest in **symbolic logic and mathematical philosophy**; . . . in principle one should be able to communicate in any symbolic logic, provided that the machine were given instruction tables which would enable it to **interpret that logical system**.

2. Turing on machine learning

> Let us suppose we have set up a machine with certain initial instruction tables, so constructed that these tables might on occasion, if good reason arose, **modify those tables**. One can imagine

> that after the machine had been operating for some time the instructions would have been altered out of all recognition, but nevertheless still be such that one would have to admit that the machine was still doing very worthwhile calculations. Possibly it might still be getting results of the type desired when the machine was first set up, but in a much more efficient manner. In such a case one would have to admit that the progress of the machine had not been foreseen when its original instructions were put in. It would be like a pupil who had learnt much from his master, but had **added much more by his own work**.

3. Turing on cognitive compatibility

> No man adds very much to the body of knowledge; why should we expect more of a machine? Putting the same point differently, the machine must be allowed to have **contact with human beings** in order that it may **adapt itself to their standards**.

AI's inventory of fundamental ideas due to Turing would not be complete without the proposal which he put forward three years later in the philosophical journal *Mind*, known today as the "Turing Test". The key move was to define intelligence *operationally*, i.e., in terms of the computer's ability, tested over a typewriter link, to sustain a simulation of an intelligent human when subjected to questioning. Published accounts usually overstate the scope proposed by Turing for his "imitation game", presenting the aim of the machine's side as successful deceit of the interrogator throughout a lengthy dialogue. But Turing's original imitation game asked only for a rather weak level of success over a relatively short period of interrogation.

> I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification [as between human and computer] after five minutes of questioning.

Presumably the interrogator has no more than $2\frac{1}{2}$ minutes of question-putting to bestow on each of the remote candidates, whose replies are not time-limited. We have to remind ourselves that—in spite of subsequent misstatements, repeated and amplified in a recent contribution to *Mind* by Robert French [9]—the question which Turing wished to place beyond reasonable dispute was *not* whether a machine might think at the level of an intelligent human. His proposal was for a test of whether a machine could be said to think at all.

Solipsism and the charmed circle

Notwithstanding the above, French's paper has important things to say concerning the role played by "subcognitive" processes in intelligent thought. He points out that "... any sufficiently broad set of questions making up a Turing Test would necessarily contain questions that rely on subcognitive associations for their answers."

The scientific study of cognition has shown that some thought processes are intimately bound up with consciousness while others take place subliminally. Further, as French reminds us, the two are interdependent. Yet other contributors to the machine intelligence discussion often imply a necessary association between consciousness and all forms of intelligence, as a basis for claiming that a computer program could not exhibit intelligence of any kind or degree. Thus John R. Searle [24] recently renewed his celebrated "Chinese Room" argument against the possibility of designing a program that, when run on a suitable computer, would show evidence of "thinking". After his opening question "Can a machine think?", Searle adds: "Can a machine have conscious thoughts in exactly the same sense that you and I have?" Since a computer program does nothing but shuffle symbols, so the implication goes, one cannot really credit it with the kinds of sensations, feelings, and impulses which accompany one's own thinking—e.g. the excitement of following an evidential clue, the satisfaction of following a lecturer's argument or the "Aha!" of subjective comprehension. Hence however brilliant and profound its responses in the purely intellectual sense recognised by logicians, a programmed computing system can never be truly intelligent. Intelligence would imply that suitably programmed computers can be *conscious*, whereas we "know" that they cannot be.

In his 1950 *Mind* paper Turing considered arguments of this kind, citing Jefferson's Lister Oration for 1949:

> Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants.

Jefferson's portrayal of conscious thought and feeling here compounds two aspects which are today commonly distinguished, namely on the one hand self-awareness, and, on the other, empathic awareness of others, sometimes termed "inter-subjectivity". Turing's comment on Jefferson disregards this second aspect, and addresses only the first. The relevant excerpt from the *Mind* paper is:

4. *Turing on the argument from consciousness*

> . . . according to the most extreme form of this view [that thought is impossible without consciousness] the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise according to this view the only way to know that a *man* thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe "A thinks but B does not" whilst B believes "B thinks but A does not". Instead of arguing continually over this question it is usual to have the polite convention that everyone thinks.

After a fragment of hypothetical dialogue illustrating the imitation game, Turing continues:

> . . . I think that most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position. They will then probably be willing to accept our test.

He thus contributes a fourth position to the previous list.

Position 4. *To the extent that possession of consciousness is not refutable in other people, we conventionally assume it. We should be equally ready to abandon solipsism for assessing thinking in machines.*

Turing did not suggest that consciousness is irrelevant to thought, nor that its mysterious nature and the confusions of educated opinion on the subject should be ignored. His point was simply that these mysteries and confusions do not have to be resolved before we can address questions of intelligence. Then and since, it has been the AI view that purely solipsistic definitions based on subjectively observed states are not useful. But Turing certainly underestimated the potential appeal of a more subtle form of solipsism generalised to *groups* of agents so as to avoid the dilemma which he posed. Following Daniel C. Dennett [5], I term this variant the "charmed circle" argument. A relationship with the notion of inter-subjectivity, or socially shared consciousness (see Colwyn Trevarthen [25, 26]) can be brought out by revising the above-quoted passage from Turing along some such lines as "the only way by which one could be sure that a machine thinks is to be a member of a charmed circle which has accepted that machine into its ranks and can collectively feel itself thinking." It is indeed according to just such social pragmatics that this issue is likely to be routinely adjudicated in the coming era of human–computer collaborative groups (see later).

But note that in the first of the two quoted passages concerning the argument from consciousness, the concluding sentence becomes truer to human society if rephrased:

> . . . it is usual to have the polite convention that everyone *who is regarded as a person* thinks.

Turing's uncorrected wording, quite unintentionally I believe, prejudges whether the polite convention would have led Aristotle to concede powers of intelligent thought to women, or Australian settlers to Tasmanian aborigines, or nineteenth century plantation owners to their slaves. Nor, conversely, would Turing necessarily have wished to withhold the polite convention if someone asserted: "My dog is a real person, and intelligent too." In his above-cited contribution on consciousness to the *Oxford Companion to the Mind*, Daniel Dennett [5] puts the point well:

> How do creatures differ from robots, real or imagined? They are organically and biologically similar to *us*, and we are the paradigmatic conscious creatures. This similarity admits of degrees, of course, and one's intuitions about which sorts of similarity counts are probably untrustworthy. Dolphins' fishiness subtracts from our conviction, but no doubt should not. Were chimpanzees as dull as sea slugs, their facial similarity to us would no doubt nevertheless favour their inclusion in the charmed circle. If house-flies were about our size, or warm-blooded, we'd be much more confident that when we plucked off their wings they felt pain (*our* sort of pain, the kind that matters).

At the outset of his paper Turing does briefly consider what may be termed the argument from dissimilarity. He dismisses the idea of "trying to make a 'thinking machine' more human by dressing it up in artificial flesh", and commends the proposed typewriter link as side-stepping a line of criticism which he sees as pointless. He evidently did not foresee the use of similarity to define charmed circles of sufficient radius to deflect the accusation of narrow solipsism. In view of the absence from his discussion of all reference to shared aspects of conscious thought, he would probably have seen it as an evasion.

The "charmed circle" criterion disallows intelligence in any non-living vehicle—so long as one is careful to draw the circle appropriately. But the idea of inanimate intelligence is in any case a difficult one to expound, for human intelligence is after all a product of the evolution of animals. This partly explains Turing's simplifying adoption of an engineering, rather than a scientific, scenario for his purpose. But the notion of performance trials conducted by a doubting client also sorted naturally with Turing's temperamental addiction to engineering images. Although himself conspicuously weak in purely mechanical skills, he startled his associates with a flow of highly original practical innovations, home-built and usually doomed. The same addiction was also discernible at the abstract level, where his gifts were not hampered by problems of physical implementation—notably in that most unlikely of purely mathematical constructions, the Universal Turing Machine itself.

Consciousness in artifacts

Searle himself is not insensitive to the appearance of special pleading. In his earlier quoted *Scientific American* article [24] he writes:

> . . . I have not tried to show that only biologically based systems like our brains can think. Right now those are the only systems we know for a fact can think, but we might find other systems in the universe that can produce conscious thoughts, and we might even come to be able to create thinking systems artificially. I regard this issue as up for grabs.

With these words he seems to accept the possibility of conscious thought in machines. But we must remember that John Searle's objection relates only to *programmed* machines or other symbol shufflers. According to the Searle canon, (i) "thinking" implies that the system is conscious, and (ii) consciousness is not possible in a programmed device which can only shuffle symbols. What if *two* "other systems in the universe" are found with identical repertoires of intellectual behaviours, including behaviour suggestive of conscious awareness? If one of the systems is implemented entirely in circuitry and the other in software, then Searle gives the benefit of the doubt to the first as a thinking system, but withholds it from the second as not being "really" conscious. What is to be done if some laboratory of the future builds a third system by faithfully re-implementing the entire circuitry of the first as software, perhaps microcoded for speed? Not only the functionality of the first machine would be reproduced, but also its complete and detailed logic and behavioural repertoire. Yet a strict reading of Searle forces the conclusion that, although the scientifically testable attributes of conscious thought would be reconstructed in the new artifact, an intangible "something" would be lost, namely "true" consciousness. The latter *could* in principle be synthesized in circuitry (see Searle: "we might even come to be able to create thinking systems artificially"), but not in software or other forms of symbol shuffling.

There is in this a counterpoint to earlier reflections by the Nobel Prize-winning neuroscientist Sir John Eccles [8], who held, like Searle today, that a being's claim to be conscious may legitimately be ignored on grounds of knowing that being's interior mechanisms:

> We can, in principle, explain all . . . input-output performance in terms of activity of neuronal circuits; and consequently, conscious-

> ness seems to be absolutely unnecessary! . . . as neurophysiologists we simply have no use for consciousness in our attempt to explain how the nervous system works.

But for Eccles it was the "activity of neuronal circuits" rather than Searle's shuffling of symbols that permitted appearances of consciousness to be discounted. Eccles has subsequently changed his view (see Popper and Eccles [21]), which we now see as having defined the limits of neurophysiology prematurely. In the same way Searle's in-principle exclusion of possible future AI artifacts risks prematurely defining the limits of knowledge-based programming. Like Eccles, Searle produces no workable definition of consciousness.

To rescue such positions from metaphysics, we must hope that a suitable Searle Test will be forthcoming to complement Turing's.

Subarticulate thought

In earlier generations students of intelligent behaviour generally ignored the phenomenon of consciousness. Some from the behaviourist camp denied that there was anything for the term to name. Turing's impact on psychology has been to move the study of cognition towards increasingly computation-oriented models. One of the ways in which these formulations differ from those of behaviourism is in allotting an important role to those aspects of consciousness which are susceptible of investigation, including investigation *via* verbal report. But as new findings have multiplied, an awareness has grown of the complementary importance of other forms of thought and intelligence. Dennett [5] remarks as follows:

> . . . the cognitive psychologist marshals experimental evidence, models, and theories to show that people are engaged in surprisingly sophisticated reasoning processes of which they can give no introspective account at all. Not only are minds accessible to outsiders; some mental activities are more accessible to outsiders than to the very "owners" of those minds!

Although having no access to these mental activities in the sense of direct awareness of their operations, the explicitly conscious forms of mental calculation enjoy intimate and instant access to their fruits. In many cases, as illustrated below with the problem of conjecturing the pronunciation of imaginary English words, access to the fruits of subliminal cognition is a necessity, even though conscious awareness of the cognitive operations themselves is not.

The Turing Test's typewriter link with the owners of two candidate minds gives no direct access to the class of "silent" mental activity described by Dennett. In the form put forward by Turing, the Test can directly detect only those processes which are susceptible of introspective verbal report. Does this then render it obsolete as a test of intelligence in machines?

The regretful answer is "Yes". The Test's didactic clarity is today suffering erosion from developments in cognitive science and in knowledge engineering. These have already revealed two dimensions of mismatch, even before questions of inter-subjectivity and socially expressed intelligence are brought into the discussion:

(i) inability of the Test to bring into the game thought processes of kinds which humans can perform but cannot articulate, and

(ii) ability of the Test to detect and examine a particular species of thought process ("cognitive skills") in a suitably programmed machine through its self-articulations; similar access to human enactments of such processes is not possible because they are typically subarticulate.

Concerning (i), the idea that mental processes below the level of conscious and articulate thought are somehow secondary was dispatched with admirable terseness by A.N. Whitehead in 1911:

> It is a profoundly erroneous truism . . . that we would cultivate the habit of thinking what we are doing. The precise opposite is the case. Civilisation advances by extending the number of important operations which we can perform without thinking about them. [28]

Moreover, although subcognitive skills can operate independently of conscious thought, the converse is by no means obvious. Indeed the processes of thought and language are dependent on such skills at every instant. The laboriously thought-out, and articulate, responses of the beginner are successively replaced by those which are habitual and therefore effortless. Only when a skilled response is blocked by some obstacle is it necessary to "go back to first principles" and reason things out step by step.

It is understandable that Turing should have sought to separate intelligence from the complications and ambiguities of the conscious/unconscious dichotomy, and to define this mental quality in terms of communication media which are characteristic of the academic rather than, say, the medical practitioner, the craftsman, or the explorer. But his Test has paid a price for the simplification. At the very moment that a human's mastery of a skill becomes most complete, he or she becomes least capable of articulating it and hence of sustaining a convincing dialogue *via* the imitation game's remote typewriter. Consider, for example, the following.

A particularly elaborate cognitive skill is of life-and-death importance in Micronesia, whose master navigators undertake voyages of as much as 450 miles between islands in a vast expanse of which less than two tenths of one per cent is land. We read in Hutchins' contribution to Gentner and Stevens' *Mental Models* [12]:

> Inasmuch as these navigators are still practicing their art, one may well wonder why the researchers don't just ask the navigators how they do it. Researchers do ask, but it is not that simple. As is the case with any truly expert performance in any culture, the experts themselves are often unable to specify just what it is they do while they are performing. Doing the task and explaining what one is doing require quite different ways of thinking.

We thus have the paradox that a Micronesian master navigator, engaging in dialogue, let us suppose, via a typewriter link with the aid of a literate interpreter, would lose most marks precisely when examined in the domain of his special expertise, namely long-distance navigation.

Procedural infrastructure of declarative knowledge

The Turing Test seems at first sight to side-step the need to implement such expertise by only requiring the machine to display general ability rather than specialist skills, the latter being difficult for their possessors to describe explicitly. But the display of intelligence and thought is in reality profoundly dependent upon some of these. Perhaps the most conspicuous case is the ability to express oneself in one's own language. Turing seems to have assumed that by the time that a fair approach to mechanizing intelligence had been achieved, this would somehow bring with it a certain level of linguistic competence, at least in the written language. Matters have turned out differently. Partly this reflects the continuing difficulties confronting machine representation of the semantics and pragmatics (i.e., the "knowledge" components) of discourse, as opposed to the purely syntactic components. Partly it reflects the above-described inaccessibility to investigators of the laws which govern what everyone knows how to do.

In spoken discourse, we find the smooth and pervasive operation of linguistic rules by speakers who are wholly ignorant of them! Imagine, for example, that the administrator of a Turing Test were to include the following, adapted from Philip Johnson-Laird's *The Computer and the Mind* [14]:

Question: How do you pronounce the plurals of the imaginary English words: "platch", "snorp", and "brell"?

A human English speaker has little difficulty in framing a response over the typewriter link.

Answer: I pronounce them as "platchez", "snorpss", and "brellz".

The linguist Morris Halle [11] has pointed out that to form these plurals a person must use unconscious principles. According to Allen, Hunnicutt and Klatt [1], subliminal encoding of the following three rules would be sufficient:

(1) If a singular noun ends in one of the phonetic segments /s/, /z/, /sh/, /zh/, /ch/, or /j/, then add the *ez* sound.
(2) If a singular noun ends in one of the phonetic segments /f/, /k/, /p/, /t/, or /th/, then add the *ss* sound.
(3) In any other case, add the *z* sound.

What about a machine? In this particular case programmers acquainted with the phonetic laws of English "s" pluralizations might have forearmed its rule base. But what about the scores of other questions about pronunciation against which they could not conceivably have forearmed it, for the sufficient reason that phonetic knowledge of the corresponding unconscious rules has not yet been formulated? Yet a human, *any* English-speaking human capable of typewriter discourse, would by contrast shine in answering such questions.

If account is taken of similar domains of discourse ranging far from linguistics, each containing thousands of subdomains within which similar "trick questions" could be framed, the predicament of a machine facing interrogation from an adversarial insider begin to look daunting. We see in this a parable of futility for the attempt to implement intelligence solely by the declarative knowledge-based route, unsupported by skill-learning. A version of this parable, expensively enacted by the Japanese Fifth Generation, is discussed in my contribution [17] to Rolf Herken's *The Universal Turing Machine*.

Among the brain's various centres and subcentres only one, localised in the dominant (usually the left) cerebral hemisphere, has the ability to reformulate a person's own mental behaviour as linguistic descriptions. It follows that Turing's imitation game can directly sample from the human player only those thought processes accessible to this centre. Yet as Hutchins reminds us in the context of Micronesian navigation, and Johnson-Laird in the context of English pronunciation, most highly developed mental skills are of the verbally *inaccessible* kind. Their possessors cannot answer the question "What did you do in order to get that right?"

This same hard fact was recently and repeatedly driven home to technologists by the rise and fall in commercial software of the "dialogue-acquisition" school of expert systems construction. The earlier-mentioned disappointment which overtook this school, after well publicised backing from government agencies of Japan and other countries, could have been avoided by acceptance of statements to be found in any standard psychological text. In John Anderson's *Cognitive Psychology and Its Implications* [2] the chapter on development of expertise speaks of the *autonomous stage* of skill acquisition:

> Because facility in the skill increases, verbal mediation in the performance of the task often disappears at this point. In fact, the ability to verbalize knowledge of the skill can be lost altogether.

Against this background, failure of attempts to build large expert systems by "dialogue acquisition" of knowledge from the experts can hardly be seen as surprising. It is sufficient here to say that many skills are learned by means that do not require prior description, and to give a concrete example from common experience. For this I have taken from M.I. Posner's *Cognition: An Introduction* [22] an everyday instance which can easily be verified:

> If a skilled typist is asked to type the alphabet, he can do so in a few seconds and with very low probability of error, If, however, he is given a diagram of his keyboard and asked to fill in the letters in alphabet order, he finds the task difficult. It requires several minutes to perform and the likelihood of error is high. Moreover, the typist often reports that he can only obtain the visual location of some letters by trying to type the letter and then determining where his finger would be. These observations indicate that experience with typing produces a motor code which may exist in the absence of any visual code.

Imagine, then, a programming project which depends on eliciting a skilled typist's knowledge of the whereabouts on an unlabelled keyboard of the letters of the alphabet. The obvious short cut is to ask him or her to type, say, "the quick brown fox jumps over the lazy dog" and record which keys are typed in response to which symbols. But you must *ask* the domain experts what they know, says the programmer's tribal lore, not learn from what they actually do. Even in the typing domain this lore would not work very well. With the more complex and structured skills of diagnosis, forecasting, scheduling, and design, it has been even less effective. Happily there is another path to machine acquisition of cognitive skills from expert practitioners, namely: analyse what the expert does; *then* ask him what he knows. As reviewed elsewhere [18], a new craft of rule induction from recording what the expert does is now the basis of commercial operations in Britain, America, Scandinavia, continental Europe, and Japan.

Psychologists speak of "cognitive skill" when discussing intensively learned intuitive know-how. The term is misleading on two counts. First, the word "cognitive" sometimes conveys a connotation of "conscious", inappropriate when discussing intuitive processes. Second, the term carries an implication of some "deep" model encapsulating the given task's relational and causal structure. In actuality, just as calculating prodigies are commonly ignorant of logical and number-theoretical models of arithmetic, so skilled practitioners in other domains often lack developed mental models of their task environments. I follow French [9] in using the term "subcognitive" for these procedurally oriented forms of operational knowledge.

Knowledge engineers sometimes call subcognitive know-how "compiled knowledge", employing a metaphor which misdirects attention to a conjectured top-down, rather than data-driven, route of acquisition. But whichever acquisition route carries the main traffic, collectively, as remarked in the passage from Whitehead, such skills account for the preponderating part, and a growing part as our technical culture advances, of what it is to be an intelligent human. The areas of the human brain to which this silent corpus is consigned are still largely conjectural and may in large part even be subcortical. A second contrast between articulate and inarticulate cerebral functions is that between the verbally silent (usually the right) cerebral hemisphere, and the logical and articulate areas located in left hemisphere. Right-brain thinking notably involves spatial visualization and is only inarticulate in the strict sense of symbolic communication: sublinguistic means of conveying mood and intent are well developed, as also are important modalities of consciousness. For further discussion the reader is referred to Popper and Eccles [21]—see also later in this paper.

The superarticulacy phenomenon

Two dimensions were earlier identified along which the Turing Test can today be seen as mismatched to its task, even if we put aside forms of intelligence evoked in inter-agent cooperation which the Test does not address at all. The first of these dimensions reveals the imitation game's inability to detect, in the human, thought processes of kinds which humans cannot articulate. We now turn to the second flaw, namely that the Test can catch in its net thought processes which the machine agent *can* articulate, but should not if it is to simulate a human.

What is involved is the phenomenon of machine "superarticulacy". A suitably programmed computer can inductively infer the largely unconscious rules underlying an expert's skill from samples of the expert's recorded decisions. When applied to new data, the resulting knowledge-based systems are often capable of using their inductively acquired rules to justify their decisions at levels of completeness and coherence exceeding what little articulacy the largely intuitive expert can muster. In such cases a Turing Test examiner comparing responses from the human and the artificial expert can have little difficulty in identifying the machine's. For so skilled a task, no human could produce such carefully thought-out justifications! A test which can thus penalise important forms of intelligent thought must be regarded as missing its mark. In my Technology Lecture at the London Royal Society [16], I reviewed the then-available experimental results. Since that time, Ivan Bratko and colleagues [3] have announced a more far-reaching demonstration, having endowed clinical cardiology with its first certifiably complete, correct, and fully articulate, corpus of skills in electrocardiogram diagnosis, machine-derived from a logical model of the heart. In a Turing imitation game the KARDIO system would quickly give itself away by revealing explicit knowledge of matters which in the past have always been the preserve of highly trained intuition.

Of course a clever programming team could bring it about that the machine was as subarticulate about its own processes as we humans. In a cardiological version of the Turing Test, Bratko and his colleagues could enter a suitably crippled KARDIO system. They would substitute a contrived error rate in the system's clinical decisions for KARDIO's near-zero level. In place of KARDIO's impeccably knowledgeable commentaries, they might supply a generator of more patchy and incoherent, and hence more true-to-life explanations. At the trivial level of arithmetical calculation, Turing anticipated such "playing dumb" tactics.

> It is claimed that the interrogator could distinguish the machine from the man simply by setting them a number of problems in arithmetic. The machine would be unmasked because of its deadly accuracy. The reply to this is simple. The machine . . . would not attempt to give the *right* answers to the arithmetical problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator.

But at levels higher than elementary arithmetic, as exemplified by KARDIO's sophisticated blend of logical and associative reasoning, surely one should judge a test as blemished if it obliges candidates to demonstrate intelligence by concealing it!

Classical AI and right-brain intelligence

Findings from "split-brain" studies have identified the co-existence in the two cerebral hemispheres of separate and under normal conditions mutually cooperating centres of mental activity. The deconnected right hemisphere (surgically separated from the left) is relatively weak in the faculties of logic and language, and is usually incapable of verbal report. Description of it as a centre of "consciousness" may therefore be questioned. But Roger Walcott Sperry, cited by Popper and Eccles in their book *The Self and Its Brain* [21], regards it as

> a conscious system in its own right, perceiving, thinking, remembering, reasoning, willing, and emoting, all at a characteristically human level. . . . Though predominantly mute and generally inferior in all performances involving language or linguistic or mathematical reasoning, the minor hemisphere is nevertheless the superior cerebral member for certain types of tasks. If we remember that in the great majority of tests it is the disconnected left hemisphere that is superior and dominant, we can review quickly now some of the kinds of exceptional activities in which it is the minor hemisphere that excels. First, of course, as one would

> predict, these are all nonlinguistic nonmathematical functions. Largely they involve the apprehension and processing of spatial patterns, relations and transformations. They seem to be holistic and unitary rather than analytic and fragmentary, and orientational more than focal, and to involve concrete perceptual insight rather than abstract, symbolic, sequential reasoning.

One should add in particular that the right hemisphere supports the recognition and appreciation of music and pictures—both important forms of human communication and cultural transmission. In his book with Karl Popper, John Eccles refers to a case reported by Gott of surgical excision of the right hemisphere, which "occurred in a young woman who was a music major and an accomplished pianist. After the operation there was a tragic loss of her musical ability. She could not carry a tune but could still repeat correctly the words of familiar songs."

In their survey of these questions, Popper and Eccles distinguish self-consciousness, located in the left hemisphere, from other forms of consciousness. In Dialogue V, Popper says:

> . . . I can't help feeling . . . that self-consciousness is somehow a higher development of consciousness, and possibly that the right hemisphere is conscious but not self-conscious, and that the left hemisphere is both conscious and self-conscious. It is possible that the main function of the corpus callosum is, so to speak, to transfer the conscious—but not self-conscious—interpretations of the right hemisphere to the left, and of course, to transfer something in the other direction too.

The corpus callosum is the two-way connecting bundle between the cerebral hemispheres. It consists of some hundreds of millions of nerve fibres, estimated to carry a total traffic of 4×10^9 impulses per second. It is this massive high-speed highway which has been surgically interrupted in the human subjects studied under the name "split-brain".

Imagine now that Turing's Test were one day to be passed by a machine simulating the kind of intelligence which remained intact in Gott's young woman patient after her operation. It is in this kind of intelligence that AI scientists have so far demonstrated most interest and success. After seeing the imitation game mastered under the special condition that the human player (as Gott's woman patient) functions in left-brain-only mode, a critic may declare that what we *now* need is a machine that can be certificated for possession of right-brain mental skills. Such a requirement would surely overwhelm the combined efforts of the world's computing community into the indefinite future. The following reasons for intractability come to mind.

The isolated right brain is rated by Eccles as "having a status superior to that of the non-human primate brain". Without a linguistic link, a first step would need to be implementation of the kind of versatile trainability on which a shepherd relies in developing intelligent behaviour in his sheep-dog. "Special testing methods with non-verbal expression", to use words taken from Sperry, are used for clinical study of human subsymbolic thinking. Equivalent methods would need to be developed for work with artificial right-brain intelligences. Imagination reels at the thought of the prodigies of innovative mechanical and electronic engineering required to build a mobile artefact which could be programmed to compete in international sheep-dog trials, or, say, which could support the behaviour required of a dumb but capable robot footman! There is a persistent suggestion that Turing (to quote the words of an anonymous informant) "also considered a robotics-style test, not one involving 'deception' of a human examiner; rather, it involved seeing to what extent the robot could succeed in the completion of a range of tasks by acting appropriately in a natural range of environments." There is a point of historical interest here. Although I cannot confirm this account from personal recollection, it fits in with a tale which I have related elsewhere [15] of Turing's early post-war days at NPL: "Turing", his colleagues said, "is going to infest the countryside with a robot which will live on twigs and rust!" Such things are still in the far future of the art of rule-based programming. Even the proprioceptive control of a single robot limb at present constitutes an almost crushing challenge. H. McIlvaine Parsons' [20] proposals are in the same direction, and contain much of interest.

Perhaps the symbolic school of AI should consider a treaty with their neural net and connectionist colleagues. Each party might renounce its claim to the other's hemispheric area, and concentrate on the cognitive functions which lie within its competence. This said, the time is nevertheless ripe for a symbolically oriented foray into one particular silent area, namely the inductive extraction of articulate models from behavioural traces of inarticulate skills.

Consciousness and human–computer interaction

Only the dominant hemisphere is logically and linguistically equipped to respond to Turing's Test. But the other hemisphere's functions include complementary forms of thought and intelligence. To complicate matters, a person's stream of awareness becomes available to others only when that person's (left-brain) discourse draws upon selected traces of conscious experience in memory. Moreover such traces partly consist of after-the-event rationalisations and confabulations. Possibly one of the needs catered for by the editing process is mnemonic. Another may be the need to explain one's feelings and/or actions to self and others. Such possibilities are consistent with Trevarthen's [26] account in *The Oxford Companion to the Mind*:

> The brain is adapted to create and maintain human society and culture by two complementary conscious systems. Specialized mo-

> tives in the two hemispheres generate a dynamic partnership between the intuitive, on the one side, and the analytical or rational, on the other, in the mind of each person. This difference appears to have evolved in connection with the human skills of intermental cooperation and symbolic communication.

As with scientific publication from an experienced laboratory, what appears in conscious recollection seems to be not so much a panoptic video of everything that went on, but rather a terse documentary, edited to highlight and establish each main conclusion. This view of the matter has survived without serious challenge since its formulation by William James [13] just a century ago:

> We see that the mind is at every stage a theatre of simultaneous possibilities. Consciousness consists in the comparison of these with each other, the selection of some, and the suppression of others, of the rest by the reinforcing and inhibiting agency of attention. The highest and most celebrated mental products are filtered from the data chosen by the faculty below that, which mass was in turn sifted from a still larger amount of yet simpler material, and so on. The mind, in short, works on the data it received much as a sculptor works on his block of stone. In a sense, the statue stood there from eternity. But there were a thousand different ones beside it. The sculptor alone is to thank for having extricated this one from the rest.

James' image of the sculptor postulates selective data destruction. Recent laboratory studies summarized by Dennett and Kinsbourne [7] indicate a slightly different metaphor. Rather than the sculptor's block of stone, Dennett proposes the writer's "multiple drafts" as a model of conscious recollection—for a comprehensive account, see his *Consciousness Explained* [6]. As far as can be, the mind's inbuilt editor squeezes parallel streams of perception into a single "story line", *not refraining even from re-arranging temporal sequences.* The following account comes from Dennett and Kinsbourne's paper.

> *The cutaneous "rabbit"* . . . . The subject's arm rests cushioned on a table, and mechanical square-wave tappers are placed at two or three locations along the arm, up to a foot apart. A series of taps in rhythm are delivered by the tappers, e.g. 5 at the wrist followed by 2 near the elbow and then 3 more on the upper arm. The taps are delivered with interstimulus intervals between 50 and 200 msec. So a train of taps might last less than a second, or as much as two or three seconds. The astonishing effect is that the taps seem to the subjects to travel in regular sequences over equidistant points up the arm—as if a little animal were hopping along the arm. Now *how did the brain know* that after the 5 taps to the wrist, there were

> going to be some taps near the elbow? The experienced "departure" of the taps from the wrist begins with the second tap, yet in catch trials in which the later elbow taps are never delivered, all five wrist taps are felt at the wrist in the expected manner. The brain obviously cannot "know" about a tap at the elbow until after it happens. Perhaps, one might speculate, the brain delays the conscious experience until after all the taps have been "received" and then, somewhere upstream of the seat of consciousness (whatever that is), *revises* the data to fit a theory of motion, and sends the edited version on to consciousness.

We guide the compressions and filterings, and manage the unavoidable distortions of the story-building, with the aid of an intellectual tool developed expressly for the purpose, the notion of causality—absent, as Russell [23] was the first to point out, from the explicit formulations of classical physics. This notion can be seen as a cognitive device for trivializing the computations while extracting the story that we can most easily understand. It follows that cause and effect may be differently attributed by different rational observers of the same events. R.E. Ornstein [19] in his *The Psychology of Consciousness* quotes from an Islamic tale:

> "What is Fate?" Nasrudin was asked by a scholar.
> "An endless succession of intertwined events, each influencing the other."
> "That is hardly a satisfactory answer. I believe in cause and effect."
> "Very well," said the Mulla, "look at that." He pointed to a procession passing in the street.
> "That man is being taken to be hanged. Is that because someone gave him a silver piece and enabled him to buy the knife with which he committed the murder; or because someone saw him do it; or because nobody stopped him?"

Design pragmatics of intelligent awareness

To many, the incommunicable (and less testable) experiences seem even more vitally important than the communicable aspects of consciousness. But for detecting and measuring intelligence in *other* agents we have to substitute for the full concept a restricted notion of "operational awareness", i.e., the testable aspect. This notion is easier to integrate into scientific usage. Moreover, news from the market-place indicates an unexpected application in the technology of graphical user interfaces for personal computers and workstations. Developments are in train in the laboratories of a number of large corporations for animated figures to appear on the screen. These personal agents are programmed to simulate awareness and intelligent interest, articulately guiding the interactive user through the operating system. Nicholas Negroponte, who directs MIT's Media Lab, remarked in the *Byte* magazine of December 1989: "Today, the PC is driven by the desk-top metaphor, but that scenario will disappear and be replaced by a theatrical metaphor. Users literally will see on their screens little *expert agents* who will do things for them."

It seems possible, then, that the software industry may come to be a more exacting designer and taskmaster of imitation games than Turing's or any other academically conceived test of machine intelligence. If the graphically displayed figures fail to muster a sufficiently convincing show of conscious awareness, users will no doubt complain to the manufacturers that their agents do not understand them. To address customer dissatisfaction of this kind, such agents will need programs capable of signalling not only coherent attention, but intentions and feelings. Negroponte's "little expert agents" must convey the motives of a teacher, or they will fail.

Conversely, simulated agents may arouse complaints that their displayed awareness, although attentive, is stilted and socially obtuse. The dimension of *social* intelligence has been noticeably absent from most discussions. Yet present-day workstation designers envisage networking environments for multiway cooperation among members of mixed task groups. Humans and machines are expected to pool diverse specialist skills in a common attack on each problem. Social intelligence by definition escapes through the net of one-on-one testing by dialogue. A version of the imitation game which can assess this form of intelligence needs to have the examiner communicate with *teams*—teams of humans, of knowledge-based robots, and (most interestingly) teams of mixed composition. Some of the design considerations go quite deep. In assessing cooperative behaviours we have to analyse, suitably translated into the realm of robot intelligence, such intimate inter-agent skills as committee work, cooperative construction of artefacts, and the collective interaction of classmates with each other and with teachers.

For forty years AI has followed the path indicated by the original form of the Test. It is no longer premature to consider extended modalities, designed to assess creative thought, subliminal and "silent" forms of expertise, social intelligence, and the ability of one intelligent agent to teach another and in turn be taught.

The imitation game in review

Let us now look back on the relation of conscious thought to Turing's imitation game.

(i) Turing left open whether consciousness is to be assumed if a machine passes the Test. Some contemporary critics only attribute "intelligence" to conscious processes. An operational interpretation of this requirement is satisfied if the examiner can elicit from the machine *via* its typewriter no less evidence of consciousness than he or she can elicit *via* the human's typewriter. We thus substitute the weaker (but practical) criterion of "operational awareness", and ask the examiner to ensure that this at least is tested.

(ii) In addition to strategies based on an intelligent agent's deep models ("understanding", "relational knowledge") we find intrinsically different strategies based on heuristic models ("skill", "know-how"). The outward and visible operations of intelligence depend critically upon integrated support from the latter, typically unconscious, processes. Hence in preparing computing systems for an adversely administered Turing Test, developers cannot dodge attempting to implement these processes. The lack of verbal report associated with them in human experts can be circumvented in some cases by computer induction from human decision examples. In others, however, the need for exotic physical supports for input–output behaviour would present serious engineering difficulties.

(iii) Conversely, where inductive inference allows knowledge engineers to reconstruct and incorporate the needed skill-bearing rules, it becomes possible to include a facility of introspective report which not uncommonly out-articulates the human possessors of these same skills. Such "superarticulacy" reveals a potential flaw in the Turing Test, which could distinguish the machine from the human player on the paradoxical ground of the machine's *higher* apparent level of intelligent awareness.

(iv) In humans, consciousness supports the functions of communicating with others, and of predicting their responses. As a next step in user-friendly operating systems, graphically simulated "agents" endowed with pragmatic equivalents of conscious awareness are today under development by manufacturers of personal computers and workstations. At this point AI comes under pressure to consider how emotional components may be incorporated in models of intelligent communication and thought.

(v) Extensions to the Turing Test should additionally address yet more subtle forms of intelligence, such as those involved in collective problem solving by cooperating agents, and in teacher–pupil relations.

(vi) By the turn of the century, market pressures may cause the designers of workstation systems to take over from philosophers the burden of setting such goals, and of assessing the degree to which this or that system may be said to attain them.

Acknowledgement

In preparing and revising this review I was assisted by facilities at the Turing Institute, Glasgow, UK, and from Virginia Polytechnic Institute and State University during tenure of a C.C. Garvin Endowed Visiting Professorship. I owe thanks to colleagues at both these institutions, and also to the referees for their helpful criticisms. I also had the great benefit of comments on an early draft from Professor Colwyn Trevarthen and from Dr Patricia Churchland, both of whom pointed me to relevant materials and sources.

References

[1] J. Allen, M.S. Hunnicutt and D. Klatt, *From Text to Speech: The MITalk System* (Cambridge University Press, Cambridge, England, 1987).

[2] J.R. Anderson, *Cognitive Psychology and Its Implications* (Freeman, New York, 3rd ed., 1990).

[3] I. Bratko, I. Mozetic and N. Lavrac, *Kardio: A Study in Deep and Qualitative Knowledge for Expert Systems* (MIT Press, Cambridge, MA, 1989).

[4] B.E. Carpenter and R.W. Doran, eds., *A.M.Turing's Ace Report and Other Papers* (MIT Press, Cambridge, MA, 1986).

[5] D.C. Dennett, Consciousness, in: R.L. Gregory, ed., *The Oxford Companion to the Mind* (Oxford University Press, Oxford, 1987).

[6] D.C. Dennett, *Consciousness Explained* (Little, Brown & Co., Boston, MA, 1992).

[7] D.C. Dennett and M. Kinsbourne, Time and the observer: the where and when of consciousness in the brain, *Behav. Brain Sci.* (to appear).

[8] J.C. Eccles (1964), Cited in: R.W. Sperry, Consciousness and causality, in: R.L. Gregory, ed., *The Oxford Companion to the Mind* (Oxford University Press, Oxford, 1987).

[9] R.M. French, Subcognition and the limits of the Turing Test, *Mind* **99** (1990) 53–65.

[10] R.L. Gregory, *The Oxford Companion to the Mind* (Oxford University Press, Oxford, 1987).

[11] M. Halle, Knowledge unlearned and untaught: what speakers know about the sounds of their language, in: M. Halle, J. Bresnan and G.A. Miller, eds., *Linguistic Theory and Psychological Reality* (MIT Press, Cambridge, MA, 1978).

[12] E. Hutchins, Understanding Micronesian navigation, in: D. Gentner and A. Stevens, eds., *Mental Models* (Erlbaum, Hillsdale, NJ, 1983) 191–225.

[13] W. James, *The Principles of Psychology* (Dover, New York, 1950) (first published 1890).

[14] P.N. Johnson-Laird, *The Computer and the Mind* (Harvard University Press, Cambridge, MA, 1988).

[15] D. Michie, Editorial introduction to A.M. Turing's chapter "Intelligent machinery" in: B. Meltzer and D. Michie, eds., *Machine Intelligence* **5** (Edinburgh University Press, Edinburgh, Scotland, 1969).

[16] D. Michie, The superarticulacy phenomenon in the context of software manufacture, *Proc. Roy. Soc. Lond. A* **405** (1986) 185–212; also in: D. Partridge and Y. Wilks, eds., *The Foundations of Artificial Intelligence* (Cambridge University Press, Cambridge, England, 1990) 411–439.

[17] D. Michie, The Fifth Generation's unbridged gap, in: R. Herken, ed., *The Universal Turing Machine* (Oxford University Press, Oxford, 1988).

[18] D. Michie, Methodologies from machine learning in data analysis and software, *Computer J.* **34** (6) (1991) 559–565.

[19] R.E. Ornstein, *The Psychology of Consciousness* (Harcourt, Brace and Jovanovich, 1977) 96; (Freeman, New York, 1st ed., 1972).

[20] H.M. Parsons, Turing on the Turing Test, in: W. Karwowski and M. Rahimi, eds., *Ergonomics of Hybrid Automated Systems II* (Elsevier Science Publishers, Amsterdam, 1990).

[21] K.R. Popper and J.C. Eccles, *The Self and Its Brain* (Routledge and Kegan Paul, London, 1977).

[22] M.I. Posner, *Cognition: An Introduction* (Scott, Foresman, Glenview, IL, 1973).

[23] B. Russell, On the notion of cause, *Proc. Aristotelian Soc.* **13** (1913) 1–26.

[24] J.R. Searle, Is the brain's mind a computer program? *Sci. Am.* **262** (1990) 20–25.

[25] C. Trevarthen, The tasks of consciousness: how could the brain do them? *Brain and Mind*, Ciba Foundation Series 69 (New Series) (Excerpta Medical/Elsevier North-Holland, Amsterdam, 1979).

[26] C. Trevarthen, Split-brain and the mind, in: R.L. Gregory, ed., *The Oxford Companion to the Mind* (Oxford University Press, Oxford, 1987).

[27] A.M. Turing, Computing machinery and intelligence, *Mind* **59** (1950) 433–460.

[28] A.N. Whitehead, *An Introduction to Mathematics* (1911).