CAN MACHINES THINK? DEEP BLUE AND BEYOND[1]

Daniel C. Dennett[2]

Medford, USA

ABSTRACT

The question "Can Machines Think?" has always intrigued researchers and philosophers. Last year the question revived when DEEP BLUE played Kasparov, but Kasparov then held his ground. However, recently DEEP BLUE defeated the World Champion, and so the question became really acute.

This article attempts to define a suitable sword-in-the-stone test, equally valid for computers and humanity. The following tests are successively discussed: the Chess Test, the Turing Test, and the Gödel Test. The Chess Test is considered to be obsolete after Kasparov's defeat. The Turing Test is illustrated by providing background on the Loebner Prize Competitions. Its restricted version is a problematic sword-in-the-stone test; its unrestricted version is certainly not a test that any machine is going to pass any time in the foreseeable future. Finally, the Gödel Test is discussed.

1. TELLING COMPETITIONS

Machines have gradually replaced human beings in many arenas. The steam drill beat John Henry, who died with the hammer in his hand, according to the famous American folksong, but that was not a contest our species needed to win, in retrospect. As we lose ground in contest after contest, is it important that there always be Some Terrific Thing that no computer (or other machine) could ever best us at?

The triumph of DEEP BLUE over Garry Kasparov last May has given the world at least one significant benefit. It has helped to bring this anxiety into focus by giving everybody a fixed point upon which they can all agree. Behind the din of those crowing "We told you so!" on the one hand and those sneering "So what?" on the other, the quiet fact remains that while the experts were not surprised – some had been predicting the victory with quiet confidence for years (*cf.* Levy, 1997) – quite a large proportion of the public, even the sophisticated public, were clearly rocked back on their heels. At least now we all know that it is indeed possible; it has been done. In a world where most of the important changes creep up on us gradually, it is convenient for us to agree upon some more or less conventional watershed items: flying across the Atlantic alone counts for more than flying the same distance over land, setting foot on the moon counts for more than sending an apparatus to the moon. Beating the World Champion at chess is a handsome example of this genre, but now that this milestone has been passed, people are wondering if it really mattered after all. Perhaps, they think, there are more telling competitions.

2. A SWORD-IN-THE-STONE TEST

In the legend of King Arthur, there is a sword stuck fast in a stone; whoever can pull out the sword is the rightful King of Britain. As a test, it has the important feature that everybody can readily see whether or not

[1] This is an extended and amended version of a lecture delivered on October 8, 1997, in Maastricht, The Netherlands. The lecture was given in the framework of the Dr. J. Tans lectures, a series of annual lectures named after one of the founders of the Universiteit Maastricht. We kindly acknowledge the preparedness of Studium Generale Maastricht to allow an (enriched) reproduction of the lecture in the *ICCA Journal*. The lecture has been published as booklet under the same title, with ISBN 90 80338524.

[2] Tufts University, Centre of Cognitive Studies, Medford, MA 02155-7059, USA. Email: ddennett@diamond.tufts.edu.

you can do it. Either you can, or you cannot. There is no need to check the photo finish or await the totalling of the judges' decisions for style points. Like the swish of the net in basketball, it provides an objective, easily read sign of success. People have been trying to define a suitable sword-in-the-stone test for humanity for hundreds of years. The pressing question reads: "is there some sword-in-the-stone test – something a human being could do that no machine could do?"

In his *Discours de la méthode*, René Descartes (1637; see also 1967) asked himself how one could tell a genuine human being from any machine. In his days, there were some rather marvelous automata, mainly clockwork devices. They were parlor tricks. Yet the question arose whether you could always tell one of these from a human being. Descartes (1637, part 5) came up with "two very certain means".

> "The first is that they [the machines] could never use words, or put together other signs, as we do in order to declare our thoughts to others. For we can certainly conceive of a machine so constructed that it utters words, and even utters words which correspond to bodily actions causing a change in its organs (e.g., if you touch it in one spot it asks what you want of it, if you touch it in another it cries out that you are hurting it, and so on). But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do. Secondly, even though such machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they were acting not through understanding but only from the disposition of their organs. For whereas reason is a universal instrument which can be used in all kinds of situations, these organs need some particular disposition for each particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act."

This is a very compelling, well-argued statement. Given when it was written and composed, I think we can agree that Descartes had every reason to think it was watertight.

Alan Turing (1950) asked himself the same question, and came up with just the same acid test – somewhat more rigorously described: it is what he called the imitation game, and we now call the Turing Test. Put two contestants, one human, one a computer, in boxes (in effect) and conduct conversations with each contestant; if the computer can convince you it is the human being, it wins the imitation game. Turing's verdict (1950, p. 442), however, was strikingly different from Descartes'.

> "I believe that in about fifty years' time it will be possible to program 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 a 70 percent chance of making the right identification after five minutes of questioning. The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."

Turing has already been proven right, I think, about his last prophecy: "the use of words and general educated opinion", as he says, has already "altered so much" that one *can* speak of machines thinking without expecting to be contradicted – "on general principles". Descartes found the notion of a thinking machine "inconceivable" and even if, as many today believe, no machine will ever succeed in passing the Turing Test, almost no one today would claim that the very idea is inconceivable. No doubt this sea change in public opinion has been helped along by the computer's progress on other feats, such as playing chess. In 1957, Herbert Simon, one of the founding fathers of Artificial Intelligence, predicted: "Within ten years a computer will be the World Champion unless the rules bar it from competition" (Simon and Newell, 1958), in other words, by 1967 a computer would be the World Chess Champion, a classic case of over-optimism, as it turns out. In August 1968, John McCarthy and Donald Michie made substantial wagers with the British chess expert, David Levy, betting that a computer would beat *him* within ten years. A year later Seymour Papert joined the original duet, adding £250 to the bet; in 1971, Ed Kozdrowicki completed the quartet. For details I refer to *More Chess and Computers* (Levy and Newborn, 1980, pp. 1-31). When their deadline came up around 1978, Levy won. Levy beat their favorite computer program, their champion. That was then, this is now. McCarthy, Michie, Papert and Kozdrowicki lost their wagers, but they were only off by a few years, not by a century or a millennium.

The reactions to DEEP BLUE's triumph have been fascinating. A few diehards have claimed that DEEP BLUE "cheated", or that Kasparov was beaten not by DEEP BLUE itself, but by a human brains trust, the designers of DEEP BLUE. This ad hoc special pleading rings hollow, at least in my ears; the same people were quite content to predict Kasparov's inevitable victory *over the computer* (not the IBM cartel) before the awesome outcome forced them to reconsider their position.

Another reaction of more than passing interest is the claim, advanced by many onlookers, that we do not learn anything from DEEP BLUE about human intelligence. The well-nigh standard dismissal claims that DEEP BLUE achieves its victories by "brute-force" computing, while Kasparov, it is claimed, uses "intuition". This is deeply misleading, since Kasparov uses brute force as well. His brain consists of over ten billion little robots – neurons – harnessed into an enormous parallel network. None of those robots understands a thing about chess, but the ensemble, magnificently orchestrated by years of training and practice, does. Kasparov's so-called intuition depends on his having a huge human brain, so it is apparently a type of brute-force computing that we do not yet understand. Kasparov himself has few if any insights into how the billions of micro-processes that compose his "intuition" do their work, but, of course, this does not show that his intuition is not "just" massive parallel computing. (It is not just the hardware, it is not just the massive parallel computing with the 10 billion little robots he has got; you have to have the software as well. I have approximately the same number of neuronal robots in my head, but I cannot do what Kasparov can do. My team of 10 billion robots cannot play chess very well at all.) While it is surely true that the particular virtual architecture of DEEP BLUE bears only faint resemblances to the virtual architecture of Kasparov's brain, no reason has yet been given to show that this is the difference of metaphysical kind that some ideologues have tried to insist on.

So what is the conclusion to draw about the Chess Test as a sort of a sword-in-the-stone test? It is a suitable sword-in-the-stone test in one regard – there is scant room for weaseling or contention over whether or not you have been checkmated – but it has been rendered obsolete by DEEP BLUE's triumph. Chess, as it now turns out, is too easy a test if you want a full-proof test only a human being can pass. It is too easy a test but it is interesting to know how many human beings over the years have thought otherwise, who thought very explicitly that no machine could possibly play chess well. This idea has had a distinguished history. Edgar Allen Poe was so certain of this that it drove him, in the nineteenth century, to unmask one of the great hoaxes of his age, von Kempelen's chess "automaton".

3. MINDS AND MACHINES

In the eighteenth century, the great Vaucanson, the French automaton builder, had made mechanical marvels that entranced the nobility, and other paying customers, by exhibiting behaviors that even today inspire our skepticism (cf. Simmen, 1968). Could Vaucanson's clockwork duck really do what it is reported to have done? "When corn was thrown down before it, the duck stretched out its neck to pick it up, swallowed, and digested it." Other ingenious artificers and tricksters had followed in Vaucanson's wake (*cf.* Chapuis and Droz, 1949), developing the art of mechanical simulacra to such a high pitch that one of them, Baron von Kempelen, an official in the court of Empress Marie Theresa of Austria, in 1769, exploited public fascination with such devices by creating a deliberate tease: a *purported* automaton that could play chess.

Von Kempelen's original machine passed into the hands of Johann Nepomuk Maelzel, who caused quite a stir in the early nineteenth century by exhibiting it as Maelzel's Chess Machine. Maelzel was an interesting character. He exhibited an automated Trumpeter of his own invention. I do not know how it worked, but it was good enough to play in public with orchestral accompaniment. He either invented the metronome or stole the invention from a Dutchman, a controversy I will leave to others to decide, but he definitely invented a musical instrument called the Parharmonicon, a sort of automated band – with flutes, clarinets, drums, trumpets, cymbals, plucked strings, violins and cellos, for which Beethoven wrote a piece of music, *Wellington's Victory*. Beethoven's famous ear trumpet was also a Maelzel artifact.

Poe was absolutely certain that Maelzel's machine concealed a human being, and his ingenious sleuthing confirmed his suspicions, which he described in detail with an appropriate air of triumph in an article in the *Southern Literary Messenger* (Poe, 1836). In fact both von Kempelen and Maelzel had employed some of the finest chess-players of their day inside the automaton, such as William Lewis, one of the great English chess masters of the nineteenth century, according to Charles Michael Carroll (1975) whose book *The Great Chess Automaton* offers a fascinating and well-researched history of this machine. At least as interesting as Poe's reasoning about how the hoax was perpetrated is his reasoning, in a letter accompanying the publication of his article, about why it *had* to be a hoax (Poe, 1836, p. 89).

> " We have never, at any time, given assent to the prevailing opinion, that human agency is not employed by Mr. Maelzel. That such agency is employed cannot be questioned, unless it may be satisfactorily demonstrated that man is capable to impart intellect to matter: for *mind* is no less requisite in the operations of the game of chess than it is in the prosecution of a chain of abstract reasoning. We recommend those, whose credulity has in his instance been taken captive by plausible appearance, and all, whether credulous or not, who admire an ingenious train of inductive reasoning, to read this article attentively: each and all must arise from its perusal convinced that a *mere machine* cannot bring into requisition the intellect which this intricate game demands ....".

So Poe thought this was an a priori truth, or at least that it was just obvious that there had to be a human being involved in the production of these chess games. We now know that however convincing this argument *used* to be, its back has been broken. It was broken in general by Darwin's theory of natural selection, which shifts the burden of proof by showing how intelligent design can arise out of mere mechanism. The particular version of the argument Poe used, with its conclusion about the impossibility of a machine playing chess, has been definitively refuted by DEEP BLUE. David Levy (1997) predicts that by the year 2003, anybody with US$ 50 will be able to buy a program running on their home computer that can beat the human World Chess Champion.

4. THE LOEBNER PRIZE COMPETITIONS

So the Chess Test has pretty well bitten the dust. Now what about the first test I mentioned to you, Descartes' test of conversation? This test has generated controversy ever since Turing proposed his nicely operationalized version of it, and has even led to a series of real, if restricted competitions. In 1991, the First Annual Loebner Prize Competition was held in Boston at the Computer Museum. Hugh Loebner, a New York manufacturer, had put up the money for a prize – a bronze medal and $100,000 – for the first computer program to pass the Turing Test fair and square. The Prize Committee, of which I was Chairman until my resignation after the third annual competition, recognized that no program on the horizon could come close to passing the unrestricted test – the only test that is of any theoretical interest as Descartes had already made clear. However, to give the computer programmers a sporting chance during the early years, some restrictions were adopted (and the award for winning the restricted test was dropped to $2000). Where Descartes and Turing had seen the importance of leaving the topic of conversation wide open, and allowing the judge complete freedom in composing the questions posed, we allowed each contestant to restrict the topic of conversation severely, and forbade the judges to ask trick questions. We also ruled that the judges had to be laypeople; those with a background in computer science were ruled ineligible.

The first year there were ten terminals, with ten judges shuffling from terminal to terminal, each spending fifteen minutes in conversation with each terminal. Six of the ten contestants were programs, four were human "confederates" behind the scenes. All the judges knew was that there was at least one computer and one human being. Each judge had to rank all ten terminals from most human to least human. The winner of the restricted test would be the computer with the highest mean rating. But of course the winner would not have to fool any of the judges, nor would fooling a judge be in itself grounds for winning; highest mean ranking was all. But just in case some program *did* fool a judge, we thought this fact should be revealed, so judges were required to draw a line across their ordered lists, separating the humans above from the machines below.

We on the Prize Committee knew the low quality of the contesting programs that first year, and it seemed obvious to us that no program would be so lucky as to fool a single judge, but on the day of the competition, I got nervous. Just to be safe, I thought, we should have some certificate prepared to award to any programmer who happened to pull off this unlikely feat. While the press and the audience were assembling for the beginning of the competition, I rushed into a back room at the Computer Museum with a member of the staff and we cobbled up a handsome certificate with the aid of a handy desk-top publisher. In the event, we had to hand out three of these certificates, for a total of seven positive misjudgements out of a possible sixty! The gullibility of the judges was simply astonishing to me. Here I was committing the sin I have so often found in others: treating a failure of imagination as an insight into necessity. But remember that in order to make the competition much easier, we had tied the judges' hands in various ways – too many ways, in my opinion. I may have chaired the committee, but I did not always succeed in persuading a majority to adopt the rules I favored. The judges had been forbidden to *probe* the contestants aggressively, to conduct conversational experiments, in effect. With this non-aggression pact in force, they sat back passively, as instructed, and let the contestants lead them, and then they were readily taken in by a thin veneer of verisimilitude which several of the programs had provided. Joseph Weinbtraub won the first competitions – indeed all the competitions in the first three years – with variations on the same program, in some ways absolutely trivial, but with a perverse and brilliant streak: It made "whimsical conversations", consisting of a large amount of canned jokes with essentially no understanding, but it adopted a clever strategy: as soon as it was asked a question, it turned around and asked the judge a question. When the judge dutifully answered it, then another question was asked, which the judge again dutifully answered, and so on. The judge fell in the trap of working very hard to give good answers to the computer, forgetting that the job was to interrogate the computer. That was Weintraub's winning strategy.

None of the seven misjudgements counted as a real case of a computer passing the unrestricted Turing Test, but they were still surprising to me. In 1991, there were still many laypeople in the Boston area who were remarkably naive about the powers of computers. The most striking manifestation of this naiveté occurred before the actual contest, during our search for suitable judges. We had advertised a job opening, discreetly, in the Boston newspapers; people with typing ability were needed to help in the administration of a contest on a particular date. We interviewed the job candidates in an office of the Cambridge Center for Behavioral Science, and each candidate was asked to prove his or her typing ability by responding for several minutes to a program running on a cheap laptop computer, sitting on a card table in front of them, not plugged in but running on battery. The program was a stripped down version of Joseph Weizenbaum's (1966) famous ELIZA program, the "doctor" that conducts the psychiatric interview by scanning for such words as "mother" and "father" and then issuing such simple canned responses as "Does your mother cause you problems?" or "Tell me more about your family." Weizenbaum's program was a good joke twenty-five years ago, and this tiny version was nowhere near as good, but it was still impressive enough to fool several job candidates. In the course of the subsequent interview, we asked candidates what they thought they had been doing during this simple audition. Several of them confidently said they had been holding a conversation with somebody, and when asked how this could possibly be true – did they think there was a tiny person inside the laptop? – they had a ready answer: there must be a cellular telephone in the computer, connecting it to their human interlocutor! Like Edgar Allen Poe more than a century earlier, they were so sure in their bones that no machine could possibly perform this simple little stunt, that they had ingeniously figured out a possible way for a human being to be at the controls.

In the second year of this competition, 1992, we uncovered another unanticipated problem with the version of the Turing Test we were administering. During the competition, as I was walking around watching what was going on, I noticed that, due to faulty briefing of the confederates, several of them gave deliberately clunky, automaton-like answers. It turned out that they were giving the silicon contestants a sporting chance by acting as they were programs! But once we had straightened out these glitches in the rules and procedures the competition worked out just as I had originally predicted: the computers stood out like sore thumbs even though there were still huge restrictions on topic. By the way, for me as a philosopher, the main lesson of the Loebner Prize Competitions was that I had underestimated the size of the gap between real experiments and thought experiments of the sorts philosophers create. The sun always shines in a thought experiment, the electricity never fails, and everything works according to plan. Moreover, you are never surprised by the results of your own thought experiment. One of the virtues of doing a real experiment is that one thereby discovers all the different background conditions left unsaid which turn out to be of non-trivial interest. It took a substantial effort to develop secure software to do this test accurately. Hundreds of questions needed to be addressed in order to protect it against fraud and unfair advantage. All this turned out to be an exercise of considerable substance, the value of which would never have been uncovered if we had left the Turing Test as a thought experiment.

In the third year, 1993, we enlisted journalists as judges, counting on their professional pride in their ability to extract the truth from those they interview to ensure a more aggressive line of questioning – while staying within the bounds of ordinary conversation. It worked, but now two of the judges made a false *negative* judgement, declaring one of the less eloquent human confederates to be a computer. On debriefing, their explanations showed just how vast the gulf was between the computer programs and the people: they had each reasoned that the competition would not have been held if there were not at least one half-way decent computer contestant, so to be on the safe side, they simply picked the least impressive human being and declared it to be a computer. But they could see the huge gap between the computers and the people as well as everybody else could. (This reasoning would not be available, for good use or bad, in the classic unrestricted Turing Test, which pits a single computer against a single human confederate; the computer can win only by being deemed more likely to be the human being.)

The Loebner Prize Competition has proven to be a fascinating social experiment, and some day I hope to write up the inside story of the early years – a tale of sometimes hilarious misadventure, bizarre characters, interesting technical challenges, and more. But it has not succeeded in attracting serious contestants from the world's best Artificial Intelligence labs. Why not? In part because passing the Turing Test is not a sensible research and development goal for serious Artificial Intelligence. It requires, in a phrase, too much Disney and not enough science. Researchers who have spent thousands of hours (and research grant dollars) developing huge, sophisticated models of language comprehension do not want to risk being beaten by a clever but simple bag of tricks cobbled together by some hobbyist in his spare time! We might have corrected that flaw by introducing into the Loebner Competition something analogous to the " school figures" in ice-skating competition: theoretically interesting (but not crowd-pleasing) technical challenges. For instance, pronouns can be ambiguous in ways that do not bother people but require deep analysis. The following example is illustrative.

> " When she walked behind his horse, he kicked her."

Who kicked whom? There are three possibilities that are consistent with the grammar, but we human beings favor one without even noticing the others. In ordinary conversation much is left unsaid, creating "enythmemes" – arguments with unstated premises. Filling in these gaps requires vast "world knowledge". Consider what is suggested but left unsaid in a simple sentence such as this:

> "She left all the presents unopened on the hall table, so she did not realize her bracelet had been returned."

Setting test passages full of such difficulties would have weeded out the tricksters. Only those programs that performed well in the school figures – the serious competition – would be permitted into the final show-off round, where they could dazzle and amuse the onlookers with some cute Disney touches. This change in the rules would have wiped out all but the most serious and dedicated of the home hobbyists, and made the Loebner Competition worth winning – and not too embarrassing to lose! When my proposals along these lines were rejected, however, I resigned from the Committee. The annual competitions continue, however, under the direction of Hugh Loebner. On the World Wide Web I recently saw the transcript of the conversation of the winning program in the 1996 competition. It was a scant improvement over 1991, still a bag of cheap tricks with no serious analysis of the meaning of the sentences. [The 7th competition in 1997 was won by the program CONVERSE, written by a team of Intelligence Research Ltd., a company headed by David Levy. CONVERSE's persona is an English-born journalist called Catherine who works as a sub-editor on a New York astrological magazine. She is British but lives in the States. The program contains 60 topics of conversation. For details see Levy (1997)[1] – Ed.] So, the Chess Test has been past recently, but I believe that for the foreseeable future, the restricted Turing Test Competitions will be at best an amusing distraction, not a serious benchmark of progress in Artificial Intelligence.

I think it is fair to say that the Loebner Prize Competition has confirmed, empirically, what people who had thought carefully about the Turing Test already suspected: it is embarrassingly easy to fool naive judges and astronomically difficult to fool expert judges. This sensitivity to variability in the judges makes the Turing Test a problematic sword-in-the-stone feat to settle the issue – though the unrestricted version is certainly not a test that any machine is going to pass, fair and square, any time in the foreseeable future. Might there, however, be a better test?

5. THE GÖDEL TEST

What about what we might call the Gödel Test? In 1931, Kurt Gödel (1931), a young mathematician at the University of Vienna, published his proof of what is now known as Gödel's Theorem, one of the most important and surprising mathematical results of the twentieth century, establishing an absolute limit on mathematical proof that is really quite shocking. Did not Gödel prove that there is something a person can do

[1] Levy, D.N.L. (1997). A Conversational Companion. *NVKI Nieuwsbrief*, Vol. 14, No. 3, pp. 68-69. ISSN 0924-4476.

that no machine can do? This argument was proposed originally by the British philosopher J.R. Lucas (1961), and was recently revived by Roger Penrose. In *The Emperor's New Mind*, and in his subsequent book, *Shadows of the Mind* Penrose (1989, 1994) argues that there is, in effect, a Gödel test which a human being can pass and no robot or computer can pass. Would this be a better test to use in the future?

What Gödel's Theorem promises the romantically inclined is a dramatic proof of the specialness of the human mind. Here is a deed, it seems to say, that a genuine human mind can perform but that no impostor, no mere algorithm-controlled robot, could ever perform. The technical details of Gödel's proof itself need not concern us; no mathematician doubts its soundness. The controversy lies in how to harness the theorem to prove anything about the nature of the mind. The weakness in any such argument must come at the crucial empirical step: the step where we *look to see* our heroes (ourselves, our mathematicians) doing the thing that the robot simply cannot do. Is the feat in question like pulling the sword from the stone, a feat that has no plausible look-alikes, or is it a feat that cannot readily (if at all) be told from mere approximations of the feat? That is the crucial question, and there has been much confusion about just what the distinguishing feat is. Below, I will very briefly give you a flavor of what the issue is.

I want to take you back to your geometry lessons in high school. Here is an example I remember from my own geometry class: Prove that all triangles enscribed in a semi-circle are right triangles. When you look at the diagram, and make a few variations, you can just sort of see that it must be true. Now the question is: can you prove it? Here we seem to see a nice contrast between just seeing with intuition that something is true, and proving it the hard way. If you are like me, you may have wondered when the teacher put a new diagram on the board whether there might be facts about plane geometry which you could *see* were true but could not prove, not in a million years. Or did it seem obvious to you that if you yourself were unable to devise a proof of some candidate geometric truth, this would just be a sign of your own personal frailty? Perhaps you thought: "There has to *be* a proof, since it is *true*, even if I myself can never find it!"

That is an intensely plausible opinion, but what Gödel proved, beyond any doubt, is that when it comes to axiomatizing simple *arithmetic* (not plane geometry), there are truths that "we can see" to be true but that can *never* be formally proven to be true. What are the truths of arithmetic? Here are some examples: 2+2=4; there is no largest prime number; numbers evenly divisible by 10 are also evenly divisible by 2. Now, how many of these truths of arithmetic exist? Infinitely many.

There are plenty of ways of deriving the truths of arithmetic from axioms, as per Euclid, but they all have an embarrassing property: they are either *inconsistent* or *incomplete.* Consider a set of axioms and ask yourself what good they are. The question is: what can you get from them? As the philosopher John Etchemendy has put it, if they are complete, you get everything you want; if they are consistent, you want everything you get! Shockingly, no set of axioms for arithmetic has both properties; either there are truths of arithmetic that you will never get, not in an eternity of cranking out proofs from the axioms, or the axioms generate falsehoods – something you definitely do not want axioms to do!

Before Gödel devised his proof, the goal of deriving *all* (and only) mathematical truth from a single set of axioms was widely regarded by mathematicians and logicians as their great project. It was considered difficult but within reach, it was like the moon landing or the Human Genome Project of the mathematics of the day. But it absolutely can*not* be done. That is what Gödel's Theorem establishes.

Now, what is the relation of this to Artificial Intelligence and the question of whether machines can think? Gödel proved his theorem some years before the invention of the electronic computer, but then Alan Turing came along and showed how it could apply to computers. It does not matter what material you make a computer out of; what matters is the program or algorithm it runs, and since every algorithm is finitely specifiable, it is possible to devise a uniform language for uniquely describing each algorithm and putting all these specifications in "alphabetical order". Turing devised just such a system, and in it, every computer – from your laptop to the grandest parallel supercomputer that will ever be built – has a unique description as what we now call a *Turing machine.* Gödel had devised a somewhat similar way of putting all *possible* axiom systems in alphabetical order, so that you could be sure that none was left out. Putting the two schemes together, Turing showed that any formal proof procedure of the sort covered by Gödel's Theorem is equivalent to one of the Turing machines, and Gödel's Theorem shows that for each such Turing machine that is a consistent prover of arithmetical truths, there is a truth it cannot prove, its "Gödel sentence". So that is what Gödel, anchored by Turing to the world of computers, tells us: every computer that is a consistent truth-of-arithmetic prover is an incomplete truth-of-arithmetic prover; it has an Achilles heel, a truth it can never prove, even if it runs till doomsday. Now we are ready for the big question: so what?

Gödel himself thought that the implication of his theorem was that human beings – at least the mathematicians among us – cannot then be just machines, because they can do things no machine could do. More pointedly, at least some part of such a human being cannot be a mere machine, or even a huge collection of gadgets. If hearts are pumping machines and lungs are air-exchanging machines, and brains are computing machines, then mathematicians' minds cannot be their brains, Gödel thought, since mathematicians' minds can do something that no mere computing machine can do.

What, exactly, can mathematicians do? This is the problem of defining the feat for the big empirical test. It will not do to say that mathematicians, unlike machines, can *prove* any truth of arithmetic, for if what we mean by "prove" is what Gödel means by "prove" in his proof, then Gödel shows that human beings – or angels, if such there be – cannot do it either (Dennett, 1978); *there is no* formal proof of a system's Gödel sentence within the system. So what is it that mathematicians can do? It is tempting to think we have already seen an example: they can do what you used to do when you looked at the blackboard in geometry class: using something like "intuition" or "judgement" or "pure understanding", they can *just see* that certain propositions of arithmetic are true. The idea would be that they do not need to rely on grubby algorithms to generate *their* mathematical knowledge, since they have a talent for grasping mathematical truth that transcends algorithmic processes altogether.

But how can we tell "genuine mathematical insight" from an impostor? How can we tell a case of somebody (or something) "grasping the truth" of a mathematical sentence from a case of somebody (or something) just wildly guessing correctly, for instance? You could train a parrot to utter "true" and "false" when various symbols were written on the blackboard in front of it; how many correct guesses without an error would the parrot have to make for us to be justified in believing that the parrot had an immaterial mind after all (or perhaps was just a human mathematician in a parrot costume)?

"Mathematicians", Penrose says, "use 'mathematical insight' to see that a certain proposition follows from the soundness of a certain system." He then goes to some length to argue that there could be no algorithm, or at any rate no practical algorithm, "for" mathematical insight. But in going to all this trouble, he overlooks the possibility that some algorithm – many different algorithms, in fact – might yield mathematical insight without being *perfect* generators of mathematical truth. And neither Roger Penrose nor anybody else can devise an empirical test that can distinguish a good-but-imperfect competence in mathematics from the ideal that Gödel shows is unreachable in any machine. Alan Turing saw all this back in 1946, anticipating Penrose's argument and refuting it in advance:

> " In other words then, if a machine is expected to be infallible, it cannot also be intelligent. There are several theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility."

6. MODELLING INTUITION

Recall the unresolved question of how Garry Kasparov uses his "intuitive" powers to play chess. Whenever we say we have solved some problem "by intuition", all that we really mean is *we do not know how* we solved it. The simplest way of modelling "intuition" in a computer is simply denying the computer program any access into its own inner workings. Whenever it solves a problem, and you ask it how it solved the problem, it should respond: "I don't know; it just came to me by intuition" (Dennett, 1969).

But could we actually construct a machine that exhibited anything even remotely resembling human intuition? In 1978, in a paper on this topic, I imagined creating a robot I called WUNDERKIND, which would go to school, like "other" children, and learn mathematics the same way they do. Eventually, in my fantasy, WUNDERKIND is taught about Turing machines and Gödel's theorem, and even learns to "hand simulate" various Turing machines, noting, along with the other children, that these Turing machines cannot prove their Gödel sentences, which all can see to be truths of arithmetic. Is this feat supposed to be beyond any robot? That is an empirical question, but Gödel's Theorem sheds no light on it. In fact, the first steps of my fantasy are coming true at MIT. I have joined Rodney Brooks and his team in the project of building a humanoid robot, named

COG, which is supposed to undergo something similar to human infancy, learning about the world the way human children do. Sending COG to school, and having COG sufficiently learn language and mathematics to stand in for WUNDERKIND, is beyond our wildest ambitions for the foreseeable future, but we may nevertheless soon learn a great deal from the COG project about how an embodied robot could begin to develop the sort of "feel" for questions that human adults exploit.

Suppose, in any case, that COG came to participate in mathematics alongside human mathematicians. And suppose we put COG in one box and a human mathematician in another box, and asked each to perform a feat of mathematical intuition. Would there be a distinguishing mark, some tell-tale balk or confusion or failure that would give away which box COG was in? Not that anybody can say. It might be true that there was an interpretation of COG as a consistent axiomatization of arithmetic, and if so, then there would be some sequence of actions COG could never undertake (unless it "broke", violating the interpretation – always a possibility). But the same might just as well be true of the human mathematician in the box. Neither COG nor any human mathematician would be *recognizable* as an axiomatic truth-of-arithmetic machine, and that is what Gödel's Theorem is about.

7. A CONCLUSION ON THINKING

The goal of defining a sword-in-the-stone feat is dubious in any case. In one sense of " machine" it is already quite clear that we are machines: we are composed of cells – there is no reason to suppose that we have any other secret ingredient – and cells are machines. A cell cannot think, in any interesting sense of " think", but a mega-machine made of a few trillion cells can think, since you and I can think, and that is what we are.

8. REFERENCES

Carroll, C.M. (1975). *The Great Chess Automaton.* Dover Publications, Inc., New York, NY. ISBN 0-486-21882

Chapuis, A. and Droz, E. (1949). Les automates. *Figures artificielles d'hommes et d'animaux. Histoire et technique.* (ed. du Griffon). Neuchâtel.

Dennett, D.C. (1969). *Content and Consciousness.* Routledge and Kegan Paul, London.

Dennett, D.C. (1978). The Abilities of Men and Machines. *Brainstorms*, pp. 256-266. Bradford Books, Montgomery, Vermont.

Descartes, R. (1637). *Discours de la méthode.* Reprinted with annotations by Gilson (1925) under the title Discourse on Method.

Descartes, R. (1967). The Philosophical Works of Descartes, Vol. 1. Translated by E.S. Haldane and G.R.T. Ross (1st ed., 1911). Cambridge University Press, Cambridge, England.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, Vol. 38, pp. 173-198.

Hofstadter, D.R. (1979). *Gödel, Escher, Bach: an Eternal Golden Braid.* The Harvester Press Limited, Stanford Terrace, Hassock, Sussex. ISBN 0 85527 757 2.

Levy, D.N.L. (1997). Crystal Balls. *ICCA Journal*, Vol. 20, No. 2, pp. 71-78.

Levy, D.N.L. and Newborn, M. (1981). *More Chess and Computers, The Microcomputer Revolution, The Challenge Match.* Computer Science Press, Inc., Potomac, Maryland. ISBN 0-914894-07-2

Lucas, J.R. (1961). Minds, Machines, and Gödel, *Philosophy*, Vol. 36, pp. 112-127.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford University Press, Oxford, England.

Penrose, R. (1994). *Shadows of the Mind. An Approach to the Missing Science of Consciousness.* Oxford University Press, Oxford, England.

Poe, E.A. (1836). Maelzel's Chess Player, *Southern Literary Messenger*. Reprinted in *Edgar Allan Poe: Essays and Reviews* (1984). Library of America, New York, NY.

Simmen, R. (ed.) (1968). *Der mechanische Mensch.* Verlag R. Simmen, Zürich. Translated in Dutch by H. Wagemans. Published by IBM Nederland N.V. and Publishers Van Lindonk, Amsterdam, The Netherlands.

Turing, A.M. (1946). *ACE Reports of 1946 and Other Papers.* (eds. B.E. Carpenter and R. W. Doran). MIT Press, Cambridge, Massachusetts.

Turing, A.M. (1950). Computing Machinery and Intelligence, *Mind, 59*, pp. 433-460.

Weizenbaum, J. (1966). ELIZA – A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM*, Vol. 9, No. 1, pp. 36-45.