united nations educational, scientific and cultural organization

international atomic energy agency

the
abdus salam
international centre for theoretical physics

H4.SMR/1330-21

" Sixth Workshop on Non-Linear Dynamics and Earthquake Prediction"

15 - 27 October 2001

Natural and Artificial Intelligence

H. Moritz

Graz Technical University
Institute of Physical Geodesy
Graz, Austria

Natural and Artificial Intelligence

Helmut Moritz

University of Technology, Graz, Austria

October 1998

The field of artificial intelligence is very fashionable today and will certainly become even more so in the future. On the one hand, we have the enormous development of computer technology; on the other hand, this area is frequently regarded with an emotional rather than rational attitude, similar to genetic technology and atomic energy. I believe, however, that in the field of artificial intelligence, we can make precise and provable statements on what is possible and where the limits are. For instance, the question of how far robots can replace men or whether they can even threaten the existence of mankind, has a clear solution given by a famous theorem published by the Austrian Kurt Goedel in 1931.

This theorem is difficult but can be explained also to the non–specialist. This will be done later. First I will start with a fairytale that goes back to the British mathematician Roger Penrose, who has written two beautiful and profound books on this subject.

The Omniscient Computer — A Fairytale

A team of scientists has built the largest supercomputer available so far, which knows practically everything and is able to perform all intelligent human activities. This wonderful machine is presented in a large auditorium to a selected public comprising eminent scientists, members of government, prominent politicians, and high–rank administrators. A television team and expert journalists are also present. The head of the scientific team invites the public to put any imaginable questions to the computer which, he says, will answer them with a human voice.

Highly intelligent questions are posed, and the computer answers them in an equally intelligent way. All the questions are on the level of university examinations, and the computer passes with honor. The admiration is general and sincere.

After one hour of examinations, all participants are too tired to put further questions. Now a small child who happens to be present, wishes also to put a question to the computer. The child says: "Dear computer, after this hard work you must be very tired, so I will put you a very simple question: How are you? How are you feeling?" The computer manages only to stammer a few disconnected words, then there is silence, suddenly one notices that something must have gone wrong, smoke rises, there is a typical smell of burnt electronics, the computer starts to glow and to melt, and finally nothing remains but a small heap of burnt debris and of molten metal. This was the end of the greatest computer of all times.

An explanation. What has happened? The computer was prepared to answer objective questions, which in fact he did in a perfect way. The computer was not prepared to answer questions which concern his own "subjectivity". The question: "How are you feeling?" is certainly not an objective question but belongs, if I may say so, to the *reflexive* questions. What does "reflexive" mean? Think of grammar: "I see a house" is an objective statement, "I see myself (in the mirror)" is a reflexive sentence. "Myself" is a reflexive pronoun, it *refers to the speaker*, in the present case to myself. We may also say that reflexivity is *self-reference*.

The computer basically can treat only objective statements; he has problems, to say the least, with reflexive statements. It thus fell into an *infinite loop* which was so strong that the computer was rapidly destroyed by overheating before it could be saved. The programmer, if prepared beforehand, would of course have had the possibility to program the computer to answer such questions evasively such as "Dear child, *such* questions are inadmissible", thus giving at least some answer. But without this, the computer was hit unprepared and reacted with a destructive infinite loop.

In principle, a computer can only work "algorithmically" or, if you wish "think algorithmically". An algorithm is the underlying computer program, which may even have some "freedom" in the form of a random number generator. As we shall see, however, man can exactly "think nonalgorithmically", which also includes "reflexive thinking" about himself.

On terminology

What, in fact, does artificial intelligence mean? Intelligence may have two meanings: (1) the property of being intelligent (in the sense of an intelligent student); (2) simply information in general. This latter sense is used in "Central Intelligence Agency" (CIA), which collects information. In the second sense, "artificial intelligence" just means "information processing by a computer". This definition, true as it is, is probably too broad and thus too trivial.

Unfortunately, the term "intelligent" is rather misused. We speak of an intelligent refrigerator which gives a signal if there is not enough milk in it. An automobile is considered intelligent if there is no sign "Fasten seatbelts" but if a soft but energetic femal voice is heard "Please fasten your seat belt". This terminology, of course, is trivial.

There are, however, nontrivial uses of the term, "*artificial intelligence*" (AI) in the scientific literature (e.g. Brookshear 1988).

Since 1945, at important universities such as MIT and Stanford University in USA, teams have been working with very ambitious goals, such as building general machines or robots that can completely imitate or replace human thinking. Optimism was great in the beginning, and big research money was available.

In 1957, a *General Problem Solver* (GPS) designed to do just this turned out to be a flop. After that, one turned to the solution of *special problems*, which are less spectacular but more important, with splendid successes at present and with even more promises for the future. Let me give some examples.

- *Language Recognition*. A text is spoken into a dictating device and the computer directly puts out the printed text, without an intermediate typist. The problem can be solved, although not yet perfectly because the context of the speech must be taken into account, and the speakers not always speak in a grammatically correct way. Still, results are surprisingly good. The same holds for automatic image processing and for computer vision.

- *Chess–playing Machines*. They have been popular from the very beginning of AI. They have little importance in themselves, but have immediately captivated human imagination. Chess is considered a very intelligent human activity, and if a man can be replaced by a machine in chess playing, then perhaps an intelligent robot can replace man in all intelligent achievements; see our fairytale at the beginning of this lecture.

  In the beginning, such machines have beaten only beginners, but they became better and better, and now even world champions do not always beat the best machines. The big advantage of the computers is their rapidity. They can compute many possible moves in advance, but they can also be programmed to incorporate the knowledge of experts and all known tricks. Thus if they can be called intelligent, the intelligence is that of excellent programmers and of human chess players whose experience has been incorporated.

- *Expert Systems* are less spectacular but much more important in practice. Consider an example, popular with fiction writers. The President of a superpower is informed that a number of rockets with atomic warheads is approaching the country and will reach it within 20 minutes. What should the president do in these 20 minutes? Calling the Security Council may be too late; at any rate it is too late for discussions. A rapid decision must be made. — An automatic decision system, an expert system, is very important. Decisions for all possible scenarios can be prepared in advance and programmed without time pressure. An expert system will provide an optimal solution for all known situations, which is available "in real time" when the danger arises and sensing devices have provided the required data.

  If such a system has be well programmed (again by *intelligent experts*!) and incorporates objective criteria for a decision, then I personally would trust such a system more than a President of unstable health or of unstable character. If there is still time to convoke the Security Council, the computer may be asked to present various options for a final decision. But be reassured: such expert systems very probably exist in reality, but still let us hope that such dangerous situations will occur only in fiction ...

More important in everyday life are medical systems. An advanced *medical expert system*, in addition to performing "logical reasoning", is also able to acquire data from measuring systems (possibly including language and image processing) and to incorporate knowledge from the enormous body of previous experience of the most prominent medical experts. If expert systems are "intelligent", the incorporated intelligence is provided by the physicians, programmers and by other *human* experts.

- *Robots* no longer belong to science fiction. They largely have replaced human workers at the assembly line for automobiles etc., relieving man of monotonously recurring simple operations. There are, however, also "intelligent robots" capable of performing surgery in cases where high precision is needed, such as replacement of hip joints or even brain surgery. The robot not only performs surgery, but also measures, processes his measuring data, and performs automatic image processing to assure that he is working at the right place within the patient. Naturally, everthing must be done under the supervision of an experienced surgeon, who can also direct or stop the robot if necessary.

  This is certainly an impressive feat of "artificial intelligence".

What computers cannot do

Is it possible that robots get more and more and more intelligent, behave more and more like human persons, and finally are tired of their subordinate role so that they make a revolution, enslaving or finally eliminating humankind?

Goedel's theorem tells us not to fear this. Robots can only think along an algorithm, however complex it may be. They have no self–consciousness: they do not know what they are doing, and cannot be made responsible for it. Thus strikes, frustration and fights of robots against mankind belong to science fiction. Now, what is Goedel's theorem?

The paradox of the liar. To understand Goedel, we must go back about 2000 years to ancient Greece, to some "philosopher" named Epimenides who discovered a logical paradox which is very easy to understand. Let me write on the blackboard the following sentence:

(L) This statement is false.

Is this statement true or false? If it is true then "This statement is false" holds, so our statement is false. If it is false, then "The statement is false" is false, so, not being false, it must be true. So our statement (L) is false if it is true, and true if it is false. Mathematicians and logicians say that this statement is true *if and only if* it is false.

Theoretically, this is an interesting paradox still much discussed in logic. It is a problem which cannot be processed, not by man and thus even less by a computer. In the computer languages which I know, this problem would put the computer into an infinite loop (yes $\rightarrow$ no $\rightarrow$ yes $\rightarrow$ no ...)

Practically, the importance of the paradox of the liar is not so great. It is of historic interest that it is already mentioned in the Bible, in St. Paul's Letter to Titus, saying that Cretans are always liars, which is testified by a Cretian philosopher himself. A big surprise for me, however, was the fact that it plays a role even in contemporary journalism. From the highly respected American news magazine TIME (August 1998, p. 36) I quote: "...revealed a notion of truth as endlessly *self-reflecting* as a fun-house mirror. It has the vertiginous feel of Epimenides' paradox which (in one version) reads, "All Cretans are liars. I am a Cretian. Therefore I am a liar." (But, of course, if I am a liar, I'm lying about being a liar, and thus I'm not.) The lies-feeding-lies *circularity* is deeply disturbing." (The italics are mine.) In writing this in an article on a recent political scandal, the journalist Ch. Krauthammer seems to possess a rare understanding of logic and philosophy.

This paradox was, in a modified form, used by Kurt Goedel, who thus discovered one of the most famous and most influential logical-mathematical theorems of all times.

Goedel's Theorem

Of all sciences, mathematics has always been the most exact. All valid mathematical theorems must, and can, be derived from a finite set of axioms. Crudely speaking, axioms are fundamental truths which are immediately recognized as correct, even self-evident. For instance, statements such as "1+1=2" or "Through two given points there passes one and only one straight line". Euclidean geometry is based on the historically first set of axioms, which were formulated already in the 3rd century B.C.

It is necessary that the axioms be *consistent*. For instance, possible axioms "1+1=2" and "1+1=1" are inconsistent. *From inconsistent axioms, all propositions, even logically contradictory ones, could be derived.* For instance, "2+2=4" and "2+2=3" could be derived as follows

To	1+1=2		To	1+1=2	
add	1+1=2		add	1+1=1	
to get	2+2=4	(true)	to get	2+2=3	(false)

This, of course, is nonsense because the axioms are inconsistent and the "axiom" "1+1=1" is manifestly false.

There are, however, more complicated instances of this general principle.

In 1931, a young mathematician and logician living in Vienna, Kurt Goedel, published a paper with the formidable title "On formally undecidable propositions of *Principia Mathematica* and related systems". The paper is extremely difficult and very few people understood its importance. Nevertheless it soon became famous among specialists.

Principia Mathematica by Russell and Whitehead is a fundamental work which claimed to furnish a complete system of axioms, by which all mathematics can be derived from logic.

What did Goedel do? He considered a proposition similar to (L) above:

(G) This statement is unprovable.

He then proved that G is derivable from the axioms if, and only if, its contrary, not–G, is also derivable! Thus, with "provable" being the same as "derivable from the axioms",

(GG) G is provable if and only if not–G is provable.

The reader will note the similarity to the paradox of the liar, discussed in the preceding section, about a proposition L: "This statement is false". We saw that L is true if and only if it is false, or in other terms,

(LL) L is true if and only if not–L is true.

Clearly, the sentence (LL) is ridiculous and pretty useless. Not so, if we consider Goedel's sentence (GG) which differs only in replacing "true" by "provable".

If G were provable, then not–G would also be provable. If a proposition is derivable together with its contrary, then the axioms of Principia Mathematica would be inconsistent. Hard to swallow, but possible.

There is, in fact, another possibility: neither G nor non–G are provable. Then (GG) would also be true because it does *not* say that G is provable, but *only* that G is provable *if* not–G is also provable. If neither G nor non–G are provable, fine.

At present it is generally assumed that the axioms of mathematics *are* consistent. Then the second alternative says that there is at least one proposition, namely G, which can never be derived from the axioms, but neither is its contrary, non–G, derivable. The proposition G is *undecidable* (see the title of Goedel's paper).

But now comes the sensation: though neither G nor non–G can be derived, it can be seen by higher–level "informal thinking" that G *must be true*. In fact, let us rephrase what we have just said:

- neither G nor non–G can be derived,
- hence, trivially, G cannot be derived,
- hence, G is unprovable.

This means that the proposition G above, which says exactly this, *must be true* (provided, of course, that our axiom system is consistent). Clearly, this proof is not a simple derivation from the axioms but involves "metamathematical" reasoning.

This proof is tricky indeed, but the reasoning, though oversimplified, is basically correct. From the darkness of undecidability there arises, at a higher level, the light of truth!

Thus there is at least *one* true proposition that cannot be derived from the axioms.

This is admittedly a somewhat difficult argument. (Never mind, Goedel's paper with all the details is even incomparably more difficult. The best "popular" presentation ist still (Nagel and Newman 1958).)

As we have seen, deduction from the axioms is a typical activity of a computer working "algorithmically" by fixed axioms and rules of deduction. The way by which G is seen to be *true* is a typical flash of *intuition*, no less rigorous than algorithmic deduction. However, this kind of rigorous intuition is typical for the human mind able to *reflect* "from a higher level" on the algorithmic work of the computer.

Perhaps a medical example can serve to illustrate the situation. A patient suffering from compulsory neurotic thinking always repeats to himself a certain argument. (It is said that an antique "philosopher" got such a compulsory neurosis by taking the antinomy of the liar too seriously, day and night repeating: L implies non–L implies L implies non–L ... Had he been able to think about this *from a higher level*, he would have recognized that this argument is really nonsense, and he would have regained his normal thinking.) In fact, one way of curing a neurotic is raising his thinking to a higher level to make him recognize the futility of such an "infinite loop" of thinking.

We have used this word purposely because also in a computer there are *infinite loops*, which must be avoided by good programming and having built–in mechanisms that stop the computer before an infinite loop occurs. Alas, all programmers know that computers nevertheless fall quite frequently into an infinite loop, and often it may be necessary to stop the computer and start it again ...

Whereas the loop (LL) is deadly but irrelevant, Goedel's formula (GG) is logically acceptable and incredibly fruitful.

Two results of Goedel's theorem should be pointed out.

1. Mathematics cannot be completely derived "algorithmically", although computer algorithms are very useful, not only in numerical computation but also in computer algebra and computer logic (e.g., theorem proving). Hopefully, mathematics is consistent; to the present day no case to the contrary seems to have been found. However, we can never be absolutely certain; an element of "Goedelian uncertainty" remains.

2. Computers working algorithmically can never be intelligent as humans are, because they cannot reflect about themselves, about their own thinking: they cannot display "creativity" or "intuition". To repeat, "intuition" in the sense used by Goedel is to recognize as true a proposition that cannot be derived from the axioms. There is nothing mystical in this, and it is as rigorous as algorithmic thinking.

Thus, computers can think only "algorithmically". Man, in addition, can think "nonalgorithmically". ("Nonalgorithmic thinking" is but another expression for "intuition" or "creativity", but it sounds less mystical.) Since computers cannot think nonalgorithmically, they can never replace human thinking.

Goedel's proof shows something which is absolutely remarkable: in contrast to a machine, man can think "at two levels": a lower level, "algorithmic thinking", is accessible to computers as well as to humans, but a higher level, "nonalgorithmic thinking", is reserved to man only.

Other terms for "nonalgorithmic thinking" are "reflexive thinking", "self-referential thinking" (remember our fairytale about the omniscient computer!), as well as "selfconsciousness" or "creativity", even "metathinking".

This thinking in two levels is not a useless hairsplitting, but the basis of Goedel's proof, which has been seen to have enormous theoretical and practical importance for artificial intelligence. By replacing "true" by "provable", Goedel has tamed the destructive energy of the paradox of the liar, turning it into a highly sophisticated logical proof (some people regard Goedel's proof as the most important achievement of mathematical logic and Goedel himself as the greatest logician of all time, with the possible exception of Aristotle).

Such a thinking "at two levels" occurs, whenever I reflect about the possible value or insignificance of my latest scientific work (I like to do this during solitary hikes). Such self-critical thinking is impossible to a computer (unfortunately also to some human persons ...). A computer will never spontaneously write on the screen "Thank you, dear programmer, your program has really been great" or "It is a shame that I must work with such a stupid program".

Multilevel thinking is quite common in philosophy. One of the most famous philosophical statements is "Cogito, ergo sum", "I think, therefore I am". This conclusion is *not* a deduction of *formal* logic, which could be done algorithmically by a computer. Instead, the conclusion follows by reflecting on the *meaning* of the fact that I am thinking, by reflecting on thinking at a higher level. This cannot be done by a computer! By the way, this is perhaps the simplest example of "nonalgorithmic thinking" and thus may help understand Goedel's argument. In fact, we may say: With Descartes, from low-level thinking or even doubting ("cogito" or "dubito"), there follows high-level certainty of existence ("sum"). With Goedel, from low-level undecidability there followed high-level truth.

Another example, perhaps less known, has also played a great role in philosophy. It is due to the Greek philosopher Plotinus (around 200 A.D.). He formulated the statement "The thinking thinks the thinking". You may say: "Of course, what else?". But try to program this statement in a computer! As far as I know, this statement cannot be formulated in any known computer language but if it could, a horribly destructive infinite loop would follow. We know the reason: a computer can work at one level only, whereas Plotinus' sentence comprises no less than three logical levels: one for the subject "The thinking", a lower level for the verb "thinks" and a still lower level for the object "the thinking".

By the way, Plotinus' theorem was very influential in philosophy: from providing the basis of St. Augustinus' theory of the Christian trinity to the dialectic triad of Fichte, Hegel and followers.

But finally back to real world. I am a strong believer in artificial intelligence, but we must respect its limits.

We "naturally" use our legs to walk from one place to another. If the two places are far apart, such as Tuzla and Sarajevo, it is better to take an automobile (if possible). The car, so to speak, is an "artificial" extension of our legs, increasing their power to cover distances. An automobile certainly is not a *replacement* for our legs, which we need even to operate the car. In the same way, *artificial intelligence is not a competitor, but an extension of our mind*, greatly increasing its power and opening new possibilities: possibilities which are fascinating indeed.

References

Brookshear J G (1988) Computer Science: An Overview, 2nd ed., Benjamin–Cummings, Reading, Mass.

Hofstadter D R (1979) Gödel, Escher, Bach: an Eternal Golden Braid, Basic Books, New York.

Moritz H (1995) Science, Mind and the Universe: An Introduction to Natural Philosophy, Wichmann, Heidelberg.

Nagel E and Newman J R (1958) Gödel's Proof, New York Univ. Press.

Penrose R (1989) The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics, Oxford Univ. Press.

Penrose R (1994) Shadows of the Mind, Oxford Univ. Press.