Strong vs. Weak AI

Matjaž Gams
Jožef Stefan Institute, Jamova 39, 61000 Ljubljana, Slovenia
Phone: +386 61 17-73-644, Fax: +386 61 161-029
E-mail: matjaz.gams@ijs.si, WWW: http://www2.ijs.si/~mezi/matjaz.html

Keywords: strong and weak AI, principle of multiple knowledge, Church's thesis, Turing machines

Edited by: Xindong Wu
Received: May 15, 1995 **Revised:** December 13, 1995 **Accepted:** December 17, 1995

An overview of recent AI turning points is presented through the strong-weak AI opposition. The strong strong and weak weak AI are rejected as being too extreme. Strong AI is refuted by several arguments, such as empirical lack of intelligence in the fastest and most complex computers. Weak AI rejects the old formalistic approach based only on computational models and endorses ideas in several directions, from neuroscience to philosophy and physics. The proposed line distinguishing strong from weak AI is set by the principle of multiple knowledge, declaring that single-model systems can not achieve intelligence. Weak AI reevaluates and upgrades several foundations of AI and computer science in general: Church's thesis and Turing machines.

1 Introduction

The purpose of this paper is to present an overview of yet another turn-around going on in the artificial intelligence (AI) community, and to propose a border between the strong (old) and weak (new) AI through the principle of multiple knowledge.

To understand current trends in artificial intelligence, the history of AI can be of great help. In particular, it records ever recurring waves of overenthusiasm and overscepticism (Michalski, Tecuci 1993):

Early Enthusiasm or Tabula Rasa Craze (1955-1965)[1]

The first AI era was impressed by the fact that human brains are several orders of magnitude slower that computers (in transmission as well as coupling speed). Therefore, making a copy of a human brain on a computer would have to result in something ingeniously better. Three subjects were predominant: (1) learning without knowledge, (2) neural modeling (self-organizing systems and decision space techniques), and (3) evolutionary learning.

Dark Ages (1965-1975)

In the second epoch it became clear that the first approach yielded no fruitful results. There were strong indications that the proposed methods were unable to make further progress beyond solving a limited number of simple tasks. After funds for artificial intelligence research were deeply cut worldwide, new approaches were searched for. This era recognized that to acquire knowledge one needs knowledge, and initiated symbolic concept acquisition.

Renaissance (1975-1980)

Research in artificial intelligence continued despite cuts in funding, since it is a subject that will probably challenge human interest forever. Taking modest aims more appropriate to the level of current technology and knowledge sometimes produced even better results than expected. The characteristics are: (1) exploration of different strategies, (2) knowledge-intensive approaches, (3) successful applications, and (4) conferences and workshops worldwide.

[1] Years are rounded by 5. Note that there are different opinions regarding the exact periods.

AI Boom (1980-1990)

Artificial intelligence R&D produced a number of commercial booms such as expert systems. Literature, conferences, funds and related events have been growing exponentially for a few years. Superprojects like the CYC project and the Fifth Generation project were in full progress approaching final stages. Artificial intelligence was reaching maturity as indicated by: (1) experimental comparisons of AI methods and systems, (2) revival of non-symbolic methods such as neural networks and evolutionary computing, (3) technology-based fields gained attention – agents and memory-based reasoning, (4) computational learning theory, (5) integrated and multistrategy systems, and (6) emphasis on practical applications. However, no generally accepted intelligent (i.e. "truly" intelligent) system was in sight.

New AI Winter (1990-1995)

Major AI projects like the Fifth Generation project or the CYC project have not resulted in intelligent or commercially successful products. Overexpectations backfired again and criticism emerged, with two basic claims:
(1) There are several indications that intelligence can not be easily achieved on digital computers with existing approaches and methodologies[2].
(2) Today's computers as well as existing approaches basically do not differ much from those of 30 years ago (apart from being faster and having better storing capacities) and, therefore, are very unlikely to approach not only human-level but also any level of intelligence established by biological intelligent systems.

Possible consequences are profound: for example, if computers can not think, then quests for true intelligence on computers are as unrealistic as searching for perpetuum mobile. Another possible implication is as follows: if computers can nevertheless think and if the brightest minds have not been able to achieve intelligence in over 30 years on the best computers available, then they must have been trying in the wrong directions.

Funds for science in general, and AI in particular are decreasing as a long-term trend.

Invisible AI plus First Dawn Approaching? (1995-...)

Invisible AI produces working systems, although it has disappeared from the first pages of scientific journals. Software engineers are adding model-based diagnoses, rule-based modules and intelligent-interface agents on top of their conventional systems. AI techniques are invisibly interwoven with existing systems. It is not top AI science, but it works.

At the same time, bold new ideas are emerging, challenging the fundamentals of computer science as well as science in general – the Turing machine paradigm, Gödel's theorem and Church's thesis.

Pollock (1989) writes: "It represents the dream of AI since its infancy, but it is a dream that has faded in much of the AI community. This is because researchers in AI have made less progress than anticipated in achieving the dream."

In the words of Minsky (1991): "the future work of mind design will not be much like what we do today".

After this short overview of AI history, the AI mega projects FGCS and CYC are analysed in Section 2. The strong vs. weak AI issue is presented in Section 3, showing the basic differences between the two approaches and describing polarisations between their proponents. The line between strong and weak AI is proposed along the principle of multiple knowledge in Section 4. The principle presents a necessary condition for better performance and true intelligence in real-life domains. Fundamentals of AI and computer science are reexamined through the weak-AI viewpoint in Section 5, including the Turing test, Church's thesis, Gödel's theorem, and Turing machines.

2 AI Mega-Projects

2.1 The Fifth Generation Computer Systems (FGCS) Project

The FGCS project (Furukawa 1993; FGCS 1993) was the first research project in Japan to embrace international collaboration and exchange (around

[2]This viewpoint is close to the one presented by Penrose (1990) – we humans would recognise any true intelligence although different from the one we possess. Of course, there would be opinions that only humans possess intelligence even in the case when an intelligent computer passed all tests. However, at present there is no such system in sight and this is only an imaginary situation.

100 scientists involved). It created a frenzy in the developed countries, fearing that Japan is going to take the lead in another central technological area – new generation computers. As a result, several other projects were started, based on logic programming (LP), the core of the FGCS project. The project was heavily based on logic programming to bridge the gap between applications and machines. Several (some concurrent) versions of Prolog (e.g. KL1) were designed to support different levels, from the user-interface to machine language. The profound effect of LP is obvious even today, as it remains one of the central areas of computer research despite recent criticism[3].

The most crucial question posed is: is logic appropriate for real-life tasks? Obviously, it has several advantages, among them a very strict formal basis, and great expressive power. However, while it may be suitable for computers and formalists, it may not be so for humans and intelligent systems in general. Arno Penzias says: "Logic is cumbersome – that's why humans rarely use it." The logical approach effectively assumes that AI is a subset of logic and that intelligence and life can be captured in a global and consistent logic form[4]. According to logicism (Birnbaum 1992)[5], knowledge representation is independent of its use – quite opposite to the new AI approach based on biological and cognitive sciences.

The progress in both logic programming and AI areas as well as in the pursuit of general-purpose parallel computers has been modest but certainly not null. Although the Fifth Generation has not been able to compete with commercial products, the rest of the world listened to it. Japan has already launched the Sixth Generation project, based on real-life domains, neural networks, optical connections, and heavy parallelism.

[3]Just recently there have been substantial cuts in LP funding in Europe.

[4]One should be careful to distinguish between different kinds of logic. Fuzzy logic, logic of informal systems, and many-valued logic seem to be quite different from the logicism analysed here. Inductive logic programming (Bratko, Muggleton 1995) is another area that should not be identified with "pure" logic approach.

[5]Note that logicism cannot be directly identified with Nilsson's work (1991).

2.2 The CYC Project

The CYC project was started by Dough Lenat in 1984 as a ten-year project (Stefik, Smoliar 1993; Lenat, Guha 1990; Lenat 1995). Substantial funding was provided by a consortium of American companies. It is based on two premises: that the time has come to encode large chunks of knowledge into a meta-system encoding common-sense knowledge, and that explicitly represented large-scale knowledge will enable a new generation of AI systems. This "knowledge is power" (the Renaissance-era slogan) approach claims that by using huge amounts of knowledge, performance and intelligence of new generation AI systems will increase substantially. The intention is to overcome one of the biggest obstacles of existing AI systems, their brittleness (dispersed isolated systems working only on carefully chosen narrow tasks).

The CYC project addresses the tremendous task of codifying a vast quantity of knowledge possessed by a typical human into a workable system. Lenat estimates (1995) that they have entered 10^6 general assertions into CYC's knowledge base, using a vocabulary with approximately 10^5 atomic terms. CYC is intended to be able to give on-line sensible answers to all sensible queries, not just those anticipated at the time of knowledge entry. Lenat and Guha estimate that this will require at least ten million appropriately organized items of information, including rules and facts that describe concepts as abstract as causality and mass, as well as specific histographic facts. CYC includes a wide range of reasoning facilities, including general deduction and analogical inference. Reasoning is done through argumentation, through comparison of pro and con arguments.

CYC is the first project of its magnitude, and therefore represents a pioneering work. Several questions and problems were posed for the first time. The whole project has strong emphasis on pragmatism – to make something workable. There are four important design characteristics: (1) the language is first-order predicate calculus with a series of second-order extensions (2) frames are the normal (general) representation for propositions, (3) nonmonotonic inferences are made only when explicitly sanctioned by the user, and (4) knowledge acquisition and inference involve different languages between which translation is

automatic.

All knowledge in CYC is encoded in the form of logical sentences, and not in diagrams, procedures, semantic nets, or neural networks. The mechanism for managing uncertainty is not as common as Bayesian networks or reason maintenance systems. One of the interesting aspects in the CYC project is the distinction between epistemological and heuristic levels of representation. A user communicates with CYC in a high level epistemological language. CYC translates queries and assertions in this language into a lower-level heuristic notation, which provides a variety of specialized inference mechanisms corresponding to special syntactic forms.

According to the authors, success will be achieved if the system works and is used by different institutions for further research and development of new (generation of) expert and knowledge-based systems[6].

There have been several strange events related to the project from the start. For example, in the overview book by Lenat and Guha (1990), there are 22 publications, of which 7 were written by the head of the project (Lenat). In (Lenat 1995) there are only 9 publications, and only 4 of them were not (co)authored by Lenat. In addition and as pointed out by one of the anonymous referees, CYC's runtime behaviour as well as the assessment of the program in (Lenat, Guha 1990) is far too brief to be convincing.

Reviewers of the project (Stefik, Smoliar 1993) generally claim that it has not succeeded to the point proclaimed by the authors (although the project is not fully completed and the final evaluation has not been published yet). Lenat even claimed that machines will start learning by themselves when the CYC computer system becomes operational around 1994 (Lenat 1989). In 1995, it is becoming clear that nothing like that is going to happen. According to critics like Dreyfus (MTCM 1992), the CYC system is as dull as any other program[7].

[6]Authors of the project have changed success criteria and basic aims a couple of times during the last ten years, obviously trying to please public interest and accommodate scientific remarks. One of these "commercial" moves was quite probably the astronomic price of the CYC system.

[7]The anonymous referees of the paper seem to share the opinion that the paper could be even more critical of the project.

On the other hand, important new understandings were arrived at, some positive and some negative, which could be very useful for new projects. In Lenat's words (1995): (CYC) " is not a bumb on a log. It saddens me how few software-related projects I can say that about these days."[8]

2.3 CYC and FGCS – AI Dinosaurs?

The two projects have addressed several fundamental questions and come with modest and in some areas even with reasonable success. CYC has managed to encode a huge amount of knowledge and the Fifth Generation project resulted in tens of working computer systems (software plus hardware). Implemented systems have worked better than commercial ones on specific tasks. Their apparent commercial failure lies in the fact that commercial computer products such as new PC's and workstations are not only more general and applicable than the products of these huge R&D projects, but also the pace of their progress was and still is faster.

Being a pioneer has its dangers, yet one has to do it if we are to get anywhere. After all, AI is constantly changing in search for true discoveries, and in a great majority of questionnaires it is predicted a great future.

But in the eyes of public, both CYC and the Fifth Generation project have not fulfilled their promises. The relative failure revived the old hypotheses that classical symbolic AI may not be able to achieve intelligence on digital computers. In the words of Dreyfus (MTCM 1992): (classical symbolic) "AI is finished".

The analogy with dinosaurs lies in the fact that CYC and FGCS represent dominant approaches and achievements of the time, but their evolutionary line is at best shaky. "Hairy", weak AI systems will probably supplement formal ones.

In the author's opinion, basic research directions in the two projects mentioned could not produce intelligent systems at all. Both projects have adopted the computationally strong-AI approach instead of at least combining it with others, e.g. cognitive weak-AI. Both projects relied on a one-sided approach, disregarding the "new school of

[8]In my personal opinion, CYC has shown that common-sense knowledge is essential for any intelligent program. That brittle systems still dominating AI are not related to any true intelligence.

AI". This new approach claims that to design an intelligent system, one has to give it all properties of intelligent creatures: unity (i.e. multiple knowledge and multistrategy approach), intentionality, consciousness and autonomy along with generality and adaptability. However, doing this will be much more difficult than previously expected.

3 Strong and Weak AI

3.1 Description

The terms "weak" and "strong" AI were originally defined by Searle (1982); here, we shall introduce similar ones based on our viewpoints.

There are several terms attached to the old and still dominant AI: symbolic, classical, formalistic, and strong. The latest alludes to several versions of the strong AI thesis. More or less they all claim that it is possible to obtain intelligence by pure algorithmic processes regardless of technology or architecture.

By weak AI we denote:

- the negation of the strong AI thesis
- adopting knowledge from interdisciplinary sciences to upgrade the computational approach.

The extreme version of strong AI is termed strong strong AI, and the extreme version of weak AI weak weak AI. Whereas strong strong AI claims that even thermostats have feelings, weak weak AI claims that only humans can have feelings because they are the only beings with souls. Both extremes fall out of the scope of this paper.

There are several analyses of the strong-weak relations. Here, we present Sloman's gradations of the strong–weak scale (Sloman 1992). His vision of weak AI is based on architectural upgrades of Turing machines. In that sense he tries to avoid mentalism and cognitive sciences completely. Instead, he tries to upgrade the formalistic Turing-machines approach with engineering knowledge.

Sloman denotes the strongest thesis of AI as T_1. Each version T_n declares something about an Undiscovered Algorithm of Intelligence (UAI). T_1 is the strongest version, claiming that every instantiation of UAI has mental abilities – all that matters are data and algorithms – no time, rich execution mechanisms, meaning. However, abstract and statical structures can not have mental abilities. An often quoted example is the book of Einstein's brain. Supposedly, this book is no different than all the information and algorithms stored in Einstein's head. Indeed, hardly anybody would claim that any book itself – be it of Einstein's brain, Turing machines or anything else – is capable of thinking or speaking. A book on its own without any execution mechanism can not perform any action at all.

A slightly modified version of T_1 is T_{1a}: every time-instantiation of UAI has mental abilities. This eliminates the book case, but has other obvious flaws. For example, if we throw a bunch of paper sheets into the air we certainly do not get anything intelligent even in the case that by chance a new interesting story emerges. The execution mechanism must be in some sort of stronger causal relation. What about Searle's Chinese room? According to Sloman the causal relation between a book (formal syntactic structures) and Searle (the execution mechanism) is too weak. There can be no understanding and intelligence in such a loose connection.

T_2 is a further modified version, requiring sufficient reliable links between program and process. This is not a strong, but a vague, mild version. Sloman analyses the properties of links between program and process from the engineering point of view. In his view, one algorithm executed on a single processor can not emulate intelligence. The process must consist of many interleaving and intensively communicating subprocesses. The architecture of the Turing machine with one algorithm and one processor (executioner) can not provide intelligence.

The difference between physical (T_4) and virtual (T_3) parallelism is similar to that between one- and many-processor architectures. One algorithm, however complicated, is not sufficient for intelligence. Parallelism has to be at the same time fine- and coarse-grained. Minsky, Moravec and Sloman have presented various parallel architectures.

Parallelism is discussed in greater detail: T_{p1} enables intelligence with a simulated continuous environment. T_{p2} needs a serial processor with time-sharing. T_{p3} states that intelligent properties can be obtained through an appropriate ne-

twork of computers.

What if any machine relying on digital technology is incapable of reproducing intelligence? T_5 declares that at least in some subsystems supercomputing power is necessary, e.g. chemistry or biology. According to Sloman, even such discovery could be very valuable for focusing further research in AI.

T_1: abstract and statical procedures can reproduce mind
T_{1a}: time instantiation of T_1 can have mental abilities
T_2: links between programs and mechanisms
T_3: virtual parallelism
T_4: physical parallelism
T_5: super-computing powers

Figure 1: Sloman's strong (top) – weak (bottom) AI scale.

In Figure 1 we can see Sloman's gradation of the strong-weak AI paradigms.

There are several other directions of weak AI indicating that the new discipline is intensively searching for new discoveries. The general approach seems promising, yet it is not clear in which particular direction the discovery of true intelligence lies. For the time being it seems that new AI is strongly related to interdisciplinary sciences, especially biological and cognitive sciences. In the words of Edelman (1992): "Cognitive science is an interdisciplinary effort drawing on psychology, computer science and artificial intelligence, aspects of neurobiology and linguistics, and philosophy."

3.2 Strong vs. Weak AI

The strong AI thesis has been attacked by Dreyfus (1979), Searle (1982), Winograd (1991) , and Penrose (Penrose 1989; 1990; 1994). According to Sloman (1992), some practitioners of AI believe in the strong strong thesis. But that is a reason for criticising them, not AI. In any field there are the "naive, ill-informed, over-enthusiastic", according to Sloman. In Sloman's opinion, the main reason for such thinking is lack of appropriate training in philosophy.

Fair to say, the author of this paper was not much different a couple of years ago. After all, all students in computer sciences get acquainted with Church's thesis and Turing machines. After a while technical details fade away, and we are left with a frame in our memory declaring that anything that can be computed is executable by the Turing machine. And that it has been shown that the proof that the Turing machine can not solve "normal" (computable) problems cannot itself be computable (operational).

Since weak AI opposes the core of not only predominant AI but also some interpretations of postulates of computer science in general, it is of no great surprise that it has been successfully suppressed until recent years. The ideas of Winograd, Dreyfus or Searle were more or less rejected in the natural and engineering sciences community. But the discussion is becoming less and less one-sided in recent years.

One of the turn-arounds was a discussion regarding the Oxford professor Roger Penrose. He is one of the most famous mathematical physicists, with several discoveries from physics (e.g. regarding black holes with Hawking) and mathematics (e.g. how to tile a plane non-periodically with only two shapes). He wrote his first book "The Emperors New Mind: Concerning Computers, Minds, and the Laws of Physics" (1989) because he was astonished by a TV debate with strong AI supporters. The title of the book alludes to the emperor's invisible dress – everybody admires it, yet there is nothing to be seen. According to Martin Gardner's foreword, Penrose is "the child sitting in the third row, a distance back from the leaders of AI, who dares to suggest that the emperors of strong AI have no clothes."

In 1994, Joseph R. Abrahamson describes Penrose as one of those "who in the name of any one of a number of gods want to destroy rationality and science. It is important to be particularly aware when one of our attempts, in however subtle a manor, to suggest this magic should supplant or even be used to embellish reason and logic."

Based on old literature citations in Penrose's book, the predominantly strong AI community harshly attacked Penrose because of his obvious lack of knowledge of current AI activities. Even more, Penrose's arguments remain debatable even inside the weak AI community.

Yet, the criticism of classical AI failed. In a reply to Abrahamson's critique, Cronin (1994)

writes: (the old) "AI community has become an arcane, closed-minded, and theoretically incestuous field of computer science." Such words certainly did not encourage friendliness between the so antagonized communities; however, they might contain at least a grain of truth especially regarding close-mindedness[9]. Angell (1993, p.15) writes: "Do those AI people really think they can capture meaning with a logico-mathematical analysis?"

In a reply to Cronin, Abrahamson (1994b) softens his criteria, posing the limit at rejecting nonscientific approaches. In this way he does not directly reject mild versions of weak AI.

There are several well-established researchers in weak AI representing the major human factor why this new wave of weak AI was not rejected as before:

- Francis Crick is probably one of the most well-deserved researchers for introducing consciousness as a legitimate subject of science. He shared a Nobel Prize for the discovery of DNA's structure in 1953. As a neuroscientist, he wants to study consciousness through the brain's internal structure.

- Another Nobel Prize winner in weak AI is Gerald M. Edelman. He shared the prize in 1972 for research on antibodies. He is the author of neural Darwinism, a theory promoting competition between groups of neurons as the basis of awareness and consciousness.

- Brian D. Josephson won his Nobel Prize in 1973 for a special quantum effect (Josephson's junction). He proposes a unified field theory encapsulating mystical and psychic experiences.

- Maurice W. Wilkes is one of computer-science pioneers and the first person ever earning money for AI-related events. In the 1992 paper in Communications of ACM he presents the opinion that classical AI is getting nowhere in the last years in the sense that all computer systems today are totally unintelligent, and that according to empirical observations intelligence may be out of reach of digital computers.

4 Principle of Multiple Knowledge

In this section a line delimiting strong from weak AI is proposed, using the principle of multiple knowledge[10] (Gams, Križman 1991). The principle is seen as an attempt to define an AI analogy of the Heisenberg physical principle which divides the world of atomic particles from the world of macro particles. Previous related work is presented, e.g. in (Sloman 1992, Minsky 1987, Minsky 1991, Penrose 1994). Our work is presented in (Gams, Karba, Drobnič 1993).

Knowledge about domain properties can be utilised as a single system (model) or as two or more subsystems, each representing a different viewpoint on the same problem. Usually, each (sub)model represents at least a part of the external world.

The **'general' thesis of multiple knowledge** states (Gams, Križman 1991): in order to obtain better performance in real-life domains, it is generally better to construct and combine several models representing different viewpoints on the same problem than one model alone, if only a reasonable combination can be designed.

'Reasonable' combination means e.g. a combination designed by a human expert. 'Performance' means e.g. percentage of successfully solved tasks.

The **'strong' thesis of multiple knowledge** states that multiple semantic models are an integral and necessary part of intelligence in any machine or being.

In real-life domains a single model can not achieve as good performance as multiple models because each model tries to fit data and noise according to its own structure and therefore tries to impose its own view. During the construction phase, it is difficult to estimate which of the models has imposed the most appropriate structure for the unseen data, and different subparts of the measurement space are typically more suitable for different models. When combining or integrating

[9] It should be noted that AI and closely related fields are becoming more and more open to discussions. For example, see (Clancey 1993; Minsky 1991; Vera, Simon 1993).

[10] While the majority of sections in this paper represent an overview of the strong-weak AI relations, this section describes the author's personal opinion and contribution.

single models it is usually not too difficult to eliminate unsuccessful parts of models.

The general thesis of multiple knowledge implies that by constructing only one model it is practically impossible to achieve the same performance as by multiple models. In other words, although multiple models can be at any time (with more or less effort) transformed into one single model with the same performance as a set of models, in general it is not possible to construct such a single model in the process of learning without designing multiple models.

Integration of models after they are designed seems not only feasible but also sensible because of reduction in storage and classification time. In our experiments (Gams, Karba, Drobnič 1993), after integration a decrease in complexity and an increase in classification accuracy was observed.

4.1 Confirmations of the Theses

Attempts to confirm the theses of multiple knowledge were performed by:

- analogy with humans, e.g. expert groups performing better than single experts; analogy to the human brain, neural Darwinism; analogy with the architecture of human brain, especially regarding split-brains. A hypothesis is presented that the human race owes its success to the rise of multiplicity in their brains (Gazzaniga 1989; Crick 1994; Brazdil *et al.* 1991; Edelman 1991).
- Empirical learning, e.g. by analyses of PAC learning, which show that a combined system works better or the same as the best single system (Littlestone, Warmuth 1991); by practical measurements.
- Simulated models, indicating that in real-life domains significant improvements can be expected when combining a couple of the best systems (Gams, Bohanec, Cestnik 1994).
- Average-case formal models, indicating that in real-world domains combining has to be only a little bit better than by chance (success rate around 0.6) in order to produce improvements (Gams, Karba, Drobnič 1991).
- Related cognitive sciences, confirming similar ideas as the Principle although not presented in a technical form (Dennett 1991).
- Quantum physics, where the multiple-worlds theory (Dewitt 1973) enables computing in multiple universes (Deutch 1985; 1992) thus representing a possible theoretical background for the Principle.

One-model systems work, but are not as useful as many-model systems in real-life domains. If top performance matters, combining or integrating several systems generally seems to be advantageous regardless of additional costs in programming and computer time.

The strong version of the Principle represents one of the necessary conditions for true AI. It is neither sufficient nor the only necessary condition. However, it does substantially narrow the search space from single-model to many-model systems. For example, over 99% of all existing computer systems and most current AI orientations are based on a single model. Intelligent systems seem to have special properties, e.g. multiplicity. These systems are very rare among all the systems. It is highly unlikely that we find (construct) them when searching in the space of all possible systems without correctly assuming their special properties.

The Principle is sometimes getting accepted as "everybody-knew-it-all-the-time". Indeed, there are many similar ideas around, e.g. Minsky's multiple representations (1991) or Sloman's parallel architectures (1992). Angell (1993, p. 15) writes: "As if every word were not a pocket into which now this, now that, now several things at once have been put!" Accepting the Principle means introducing weak AI and leads to fundamental changes in future progress in AI and computer science alike.[11]

5 Fundamentals of AI and Computer Science

Weak AI reexamines and disputes the soundness of several well-established scientific fundamentals: Turing's test, Gödel's theorem, Church's thesis, and the Turing machine.

[11] According to the Principle, many research directions will not produce true intelligence, meaning that efforts, achievements and future funding in that areas are doubtful.

5.1 Turing's Test

When Turing nearly half a century ago posed his famous question "Can computers think", electronic computers were just emerging. The back-bone of his test is a detective probabilistic quiz in which an interrogator has to be sufficiently sure which of the two subjects communicating through a computer interface (terminal plus keyboard) is human and which computer, given limited time. Turing believed that his test would be passed in around 50 years when computer storage capacity reached 10^9. By then, "an average interrogator would not have more than 70 per cent chance of making the right identification (as between human and computer) after five minutes of questioning."

During years, several modifications of Turing's test have being proposed, e.g. the total Turing test (TTT) in which the subject has to perform tasks in the physical world such as moving blocks. Other remarks imply that the original test is (1) too easy since it is based on typed communication only, (2) too narrow since it is basically an imitation game, (3) too brittle since it can not reveal the internal structure of thinking processes – Searle's basic claim (Searle 1982), and (4) too difficult since no animal and many humans (e.g. handicapped) are unable to compete at all, and intelligence can be displayed well below average-human level. All these remarks have their counterarguments, e.g. that (1) communication through typing is more than relevant to evaluate the intelligence of a subject, e.g. by the IQ tests, (2) such communication allows very rich possibilities of questions and themata, (3) it is not possible to reveal the human thinking process either, and (4) if the Turing test (TT) is too difficult then the limited Turing test (LTT) can be applied. Indeed, such is the case in practical contests held annually (Shieber 1994). TT remains probabilistic, approximate, detective, fundamentalistic, behaviouristic and functional.

Although the Turing test is heavily analysed and disputed, it remains the most interesting scientific test up to date, offering important implications.

The latest analyses of the Turing test were performed by Turing's contemporary Donald Michie (1993). In his opinion, there are two obstacles an intelligent computer system has to face in order to approach passing it:

1. subarticulacy – the human inability to articulate specific activities although performed by humans, and

2. superarticulacy – the ability to explain particular thought processes in a suitably programmed machine although being subarticulate in humans.

Regarding the first point, humans can not articulate their internal thought processes, which are sometimes more transparent to observers than to themselves. Therefore, how can human knowledge be transformed into computer systems if humans are not able to specify it?

The second point poses another problem. Computer programs are by default traceable – meaning their decisions can be traced and reproduced. Even systems like neural nets or numerical procedures can be 'understood' up to a point, and simulated by other transparent systems. All computer systems, therefore, have abilities nonexistent in humans.

Some of these questions were discussed already by Turing. He proposed that machines would have to play the imitation game, thus simulating thought processes while inherently being different. While it is not yet clear whether digital machines can achieve intelligence at all, it is becoming accepted that on digital computers, systems simulating human thought processes will be essentially different from humans. In light of this conclusion, the claim of connectionists – that sufficiently complex neural networks will be effectively the same as the human brain – is hard to accept. Even if neural networks were to achieve the performance of a human brain, it would be possible to extract weights, topology and other characteristics of nets. By not being able to do it in humans, one (of many) unavoidable substantial difference appears. The "End of Innocence" period, together with empirical verification, brings new insights, displaying the naivete of existing approaches and opening new directions. The Turing test indicates substantial differences between formal machines and real-life beings.

Weak AI is in general satisfied with less than passing the Turing test. For example, artificial life and evolutionary computing try to simulate rather primitive forms of life. Brooks (1991) proposes intelligence without reasoning,

low-intelligence robots (insects) without symbolic internal representation of the external world. Sloman (1992) finds the Turing machine rather unrelated to real life. It represents an artificial machine very capable for specific formal tasks only. Sloman, Penrose and also people in general tend to believe that even animals can display certain aspects of intelligence when solving real-life problems. On the other hand, while machines can solve difficult formal problems which are often practically unsolvable even by humans and definitively unsolvable for all animals, they are still regarded as totally unintelligent.

5.2 Church's Thesis and Turing Machine

Around 1930 Church, Gödel, Kleene, Post, Turing and others tackled questions such as: what can be computed and what not, are all statements either provable or not inside a formal system? They have come with basic concepts that represent a backbone of today's computer science.

Church's thesis is the assertion that any process that is effective or algorithmic in nature defines a mathematical function. These functions form a well-defined class, denoted by terms such as recursive, λ-definable, Turing computable. All these functions are computable by the Turing machine, a formal model of computers. Anything that a digital or analog computer can compute, be it deterministic of probabilistic, is computable by the abstract Turing machine, given enough time and space. The problems that the Turing machine can not solve are unsolvable for present and future formal computer systems as well, be it simple PC's, supercomputers or parallel connectionist machines.

Church's thesis provides the essential foundation for strong AI. If computable problems are solvable by the Turing machine then digital computers can solve them if only they are quick enough. Therefore, achieving true intelligence on computers demands only very fast hardware with sufficient memory capabilities and a program. In Abrahamson's opinion (1994) it is only a matter of time and technological progress.

In general, there are two major philosophical orientations regarding the human mind and our world in general: mentalistic and mechanistic. Mechanicists regard mind as a material object obeying the laws of nature. Mind is a (biological, physical ...) machine. Mentalists see mental states as something beyond formal sciences (mild version) or even extramaterial, i.e. outside the real world (strong version). Church's thesis implies that its computational essence can not be refuted by effective means. It means that the opposing hypothesis can not be effective at all, or in other words, it can not be computed in the general meaning of the word.

The strong principle of multiple knowledge collides with the direct explanation of Church's thesis. One possible compromise is that although intelligent models can be – at least in principle, with unknown practical problems – designed and executed on any Turing machine, it is not possible to design intelligent computer programs in the form of a single model not consisting of multiple models. Therefore, if the program on the Turing machine is multiple enough and has the needed additional properties, it could simulate intelligence. However, the principle does not exclude the other possibility – that true intelligence can not be achieved on Turing machines at all, that stronger computational mechanisms having explicit multiplicity at the core of the computing process are necessary.

Practically all weak AI researchers in this or another way distance their ideas from Church's thesis (see Section 3). Neuroscientists (Edelman 1992) propose their models of the brain. Physicists propose new physical theories enabling new computing mechanisms – Penrose proposes microtubules where quantum effects in relation to the correct quantum gravity enable supercomputing powers. Deutch (1992) proposes a quantum Turing machine.

Sloman's viewpoint is similar to the principle of multiple knowledge based on the engineering architecture of the computing machine. Theoretically, it has been proven that the computational power of one Turing machine is equal to the power of many parallel machines. From the engineering point of view this is not the case. The key is not in speed or time, but in the architecture. For example, a fatal error in one processor simulating parallel computing causes malfunction in serial architectures yet is usually only a smaller obstacle in appropriate parallel hardware architectures. If one processor simulates several virtual processors

then it must constantly check the internal states of each parallel process. This disables true asynchronous interaction with complex real-life environments. Although the parallel and sequential process display equal computational powers, they substantially differ in causal relations.

5.3 Seeing the Truth of Gödel's Sentence

In his 1931 paper, Gödel showed that for any formal system F broad enough to express the arithmetic of natural numbers, there is a construction of a formula $P_k(k)$ where k is the Gödel's number of that formula itself. This well-defined formula is denoted by $G(F)$. Gödel's theorem states that if F is consistent, there can be no derivation of $G(F)$, and if F is omega-consistent, no derivation of $\neg G(F)$. Therefore, $G(F)$ is undecidable (unprovable), and the formal system F is incomplete.

Not only that Gödel's theorem is formally provable, computer programs such as SHUNYATA (Ammon 1993) have been able to automatically reproduce, i.e. rediscover the proof.

By proving his theorem Gödel demolished the strong formalistic approach in science. He proved that at least one formula (statement, sentence) can not be proven inside a formal system (later it was found that there are many such statements). Therefore, there is no way a formal machine can prove a specific sentence constructed by a formal (legal) procedure.

Many relevant researchers including Gödel and Turing thought that although the proof shows that it is not possible to formally prove $G(F)$, $G(F)$ is nevertheless true. Of course, no formal proof of $G(F)$ can be constructed inside F since it has been formally proven that such a proof does not exist. Therefore, how can $G(F)$ be seen as true by humans? In 1961 Lucas presented his view of this paradoxical situation hypothesising what happens if humans use some kind of a formal algorithm UAI. This idea was revived and extended by Penrose (1989).

Lucas proposes – in his viewpoint – a valid mathematical procedure for seeing the truth of $G(F)$. Namely, if the sentence asserts about itself that it is not provable, and the formal proof showed that $G(F)$ can not be proved, then the sentence is obviously true. Therefore, humans can see at $G(F)$ is true.

Penrose's extension is as follows: even if a human uses some kind of (probably very complex) formal algorithm UAI executable on a Turing machine, and we construct a formal Gödel's sentence $G(UAI)$ for that algorithm, he can see the truth of it. Not only Penrose and mathematicians, probably all students in natural and technical sciences can intuitively see (or have that feeling of) the truth of Penrose's line of reasoning. Therefore, we can assume that all humans are at least in principle able to see it. Furthermore, all humans use similar processes when seeing the truth of Gödel's sentence.

Since formal systems are not able to formally prove the truth of Gödel's sentence, and humans can see it, humans do not always apply formal algorithms (e.g. UAI). Therefore, since humans can in principle reproduce anything that Turing machines can, and Turing machines in principle can not reproduce all things humans can (e.g. seeing the truth of Gödel's sentence), Turing machines do not possess all computational powers that humans do. Since Turing machines are capable of reproducing any computation by digital computers, true intelligence can not be achieved on digital computers.

Among the common objections to this kind of reasoning are the following:

- it is not possible to see that $G(F)$ true since this requires proving that F is consistent[12];
- $G(F)$ can be seen to be true by flible and incomplete procedures (similar to the ones humans use);
- Gödel's theorem is not related to real life; it is just a formal matter relevant to formal systems. Although this means that we have to reject deductive semantics as means of describing human intelligence, we can endorse other types of inference, e.g. abductive logic.
- in a computationally stronger meta$F$ it is possible to formally prove a statement (theorem) $provable(metaF, G(F))$.

[12] As pointed out by Boolos, Chalmers, Davis and Perlis (Penrose 1990), the consistency of complex mathematical systems, e.g. ZF systems, can not be proved. This means that nobody, Turing machines and Penrose included, can prove or even see the truth of Gödel's sentence in ZF systems.

The most fundamental denial of Penrose's argument was presented by Sloman (1992). He attacked the core meaning of Gödel's theorem: Gödel's sentence does not mean what it seems to mean, and Penrose can not see the truth of $G(F)$ since there are models in which it is true and those in which it is false.

The first premise does not seem to be justified as shown by Bojadžiev (1995).

Sloman's claim is based on constructing two models: of $(F, G(F))$ and of $(F, \neg G(F))$. This is valid since neither $G(F)$ nor $\neg G(F)$ are provable in F, if consistent. Now, nobody can see the truth of $G(F)$ in $(F, \neg G(F))$, Penrose concluded.

However, in models of $(F, \neg G(F)$ it is possible to establish the truth of $\neg G(F)$, therefore, $G(F)$ is not unprovable anymore if $(F, \neg G(F)$ is consistent. Extended models of F usually do not correspond to classes of universal Turing machines. This is a common case in computational capabilities of systems: stronger mechanisms can often answer puzzles in weaker mechanisms, yet have their own undecidable questions. Sometimes it is even sufficient to apply meta–reasoning inside systems with the same computational powers, but again new undecidable questions can be produced. For example, it has been formally proven by a meta-system that Gödel's sentence $G(F)$ is true in natural numbers if F is consistent. Therefore, the truth of Gödel's theorem in certain mathematical models, e.g. in Peano Arithmetic can be formally proven outside F if it is consistent.

Here we shall translate the same problem into the world of Turing machines. Namely, Gödel's theorem corresponds to the halting problem of Turing machines, i.e. to the question if a Turing machine can in general predict whether a Turing machine will stop or not. It has been formally proven that the halting problem is in general undecidable (Turing 1936; Hopcroft, Ullman 1979). Furthermore, the concept of Gödel's theorem is so fundamental for formal systems that it can be reproduced in many forms (see for example Penrose's second book (1994)).

Consider for example an Algol-like procedure U which shows that a procedure can not determine whether it will stop or not. Reasoning starts with the hypothesis that there exists a procedure T which can determine for any procedure *proc* whether it stops or not. Then we construct a procedure U which includes the procedure T. If U itself (self-reference) is given as an input for U, it should stop when it should not (i.e. $T(U)$ is $false$) and vice versa. Since the transformation from T to U is legal inside the same description mechanism of Turing machines, and U cannot exist, T cannot exist. Therefore, a procedure which determines for any procedure whether it will stop or not, does not exist.

```
procedure U(proc);
begin
        while T(proc) do;
        write('OK');
end;
```

The self-referential applicability of U, and the halting problem in Turing machines and formal programming languages are beyond reasonable doubt. Furthermore, high-school students usually do not have troubles seeing or understanding the paradoxical nature of the halting problem.

Penrose replies that there is no reason for dealing with unsound or incomplete systems. Under this assumption it is possible to see the truth of Gödel's sentence, it is possible to formally prove it outside F, and quite probably possible to duplicate Penrose's semantical reasoning about truth by special meta-systems.

In summary, Penrose's version of the Gödel theorem and the halting problem represents an interesting hypothesis, however is not proven. On the other hand, several attempts to formally disprove Penrose's version have been formally proven to be wrong.

6 Discussion

The history of AI teaches us that the only constant is its ever-changing nature. In recent years new, fresh ideas are coming from interdisciplinary sciences – neurobiology, philosophy, cognitive sciences. In this way, the computational approaches are being enriched and upgraded.

Weak AI reexamines basic postulates of AI and computer science. In regard to *Turing's test*, proponents of weak AI see the test as an indicator of important differences between humans and computers. Computer systems can explain their line of reasoning in detail. Humans do not know how

reasoning is performed in their heads and do not know how to reveal (transplant) that to computers. Just passing the test is not sufficient to be accepted as intelligent. A computer chess program beating most humans is not intelligent although it performs brilliantly compared to an average human. Animals are not capable of playing chess, yet some of them show properties of intelligence while computers are regarded as totally unintelligent.

By-passing *Church's thesis*, weak AI does not accept that one *Turing machine* performing one algorithm is sufficient to achieve intelligence. The *principle of multiple knowledge* proposes multiple-model structures as one of necessary conditions for intelligent systems. Extreme viewpoints see digital computers as incapable of achieving intelligence.

There are several indications that the human brain is computationally more powerful than digital computers, e.g. observed through the progress of computer power and the lack of computer intelligence. Theoretical analyses are often performed through the *Gödel theorem* and *halting problem.*

The principle of multiple knowledge dictates a step-up of complexity from one optimal model to an optimal combination of models. It upgrades the centuries old *Occam's Razor* indicating that the Razor can be even misleading when blindly applied. However, an upgraded version of Occam's Razor might be valid in the multiple-model world. Similarly, human knowledge is seen as significantly more complex than currently expected. Multiple models introduce an additional level of combinatorial explosion, thus making knowledge less transparent, more difficult to store, and more powerful.

Clashes between strong and weak AI proponents may help sift new ideas and eliminate unsound attempts. Weak AI is still in the brainstorming state – lots of new ideas and not many confirmed achievements. Weak AI is getting accepted as another discipline researching consciousness and relations to computers. Similarly, most of new nonsymbolic approaches in AI were rejected at first and then accepted, be it neural networks or evolutionary computing.

How can weak AI be proven wrong? The simplest proof would be constructive – to design a single-model computer system capable of true intelligent behaviour. Note that just designing an intelligent computer system executable on a Turing machine is not enough.

How can strong AI be proven wrong? There are several possibilities. For example, it is enough that the Penrose's hypothesis about Gödel's theorem gets proven. Or that the principle of multiple knowledge gets proven. Or that neuroscience produces substantial new discoveries about the human brain. Or that a new physical theory gets proven. Or ...

Today, the house of science is based on empirical validation and formal verification. Formal verification is well within the domain of Turing computable functions. The fear that weak AI is attacking the core of science by reevaluating Church's thesis and other scientific postulates is not grounded. For example, if Penrose's ideas get accepted, meaning that unprovable true functions are computable by humans but not by computers, scientific knowledge will essentially expand. Science will expand even if the principle of multiple knowledge gets accepted. Instead of relying on formal models, other aspects will gain prominence, e.g. engineering or cognitive enrichments of formal sciences.

Acknowledgments

This work was supported by the Ministry of Science, Research and Technology, Republic of Slovenia and was carried out as part of different European projects. Research facilities were provided by the "Jozef Stefan" Institute. The author is indebted to the two anonymous reviewers and in particular the editor Xindong Wu for their constructive comments on the paper. Special thanks to Damjan Bojadžiev for careful reading and correcting the paper, and to Mare Bohanec, Matija Drobnič, Nada Lavrač and Donald Michie for helpful remarks.

References

[1] K. Ammon (1993), An Automatic Proof of Gödel's Incompleteness Theorem, *Artificial Intelligence*, **61**, pp. 291-307.

[2] I. O. Angell (1993), Intelligence: Logical or Biological, Viewpoint, *Communications of the ACM* **36**, pp. 15-16; 110.

[3] J. R. Abrahamson (1994), Mind, Evolution, and Computers, *AI magazine*, Spring 1994, pp. 19-22.

[4] J. R. Abrahamson (1994b), A Reply to Mind, Evolution, and Computers, *AI magazine*, Summer 1994, pp. 8-9.

[5] L. Birnbaum (1992), Rigor Mortis: A Response to Nilsson's "Logic and Artificial Intelligence", *Foundations of artificial intelligence*, pp. 57-79, (ed.) D. Kirsh, MIT/Elsevier.

[6] D. Bojadžiev (1995), Sloman's View of Gödel's Sentence, *Artificial Intelligence* **74**, pp. 389-393.

[7] I. Bratko and S. Muggleton (1995), Applications of Inductive Logic Programming, *Communications of the ACM* **38**, pp. 65-71.

[8] P. Brazdil, M. Gams, L. Sian, L. Torgo and W. van de Velde (1991), Learning in Distributed Systems and Multi-Agent Environments, *Proc. of EWSL-91*, Porto, Portugal.

[9] R. A. Brooks (1991), Intelligence without Representation, *Artificial Intelligence* **47**, pp. 139-160.

[10] W. J. Clancey (1993), Situated Action: A Neuropsychological Interpretation Response to Vera and Simon, *Cognitive Science* **17**, pp. 87-116.

[11] F. Crick (1994), *The Astonishing Hypothesis, The Scientific Search for the Soul*, New York.

[12] M. R. Cronin (1994), A Reply to Mind, Evolution, and Computers, *AI magazine*, Summer 1994, p. 6.

[13] D. C. Dennett (1991), *Consciousness Explained*, Little Brown.

[14] D. Deutch (1985), Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer, *Proceedings of Royal Society*, pp. 97-117.

[15] D. Deutch (1992), Quantum Computation, *Physics Word*, pp. 57-61.

[16] B. S. Dewitt (1973), *The Many-Worlds Interpretation of Quantum Mechanics*, Princenton University Press.

[17] H. L. Dreyfus (1979), *What Computers Can't Do*, Harper and Row.

[18] G. Edelman (1992), *Bright Air, Brilliant Fire, On the Matter of the Mind*, Penguin Books.

[19] FGCS (1993), The Fifth Generation Project, *Communications of the ACM* **36**, pp. 46-100.

[20] K. Furukawa (1993), Fifth Generation Computer Systems (FGCS) Project in Japan, *Japan Computer Quarterly* **93**, pp. 1-33.

[21] M. Gams, M. Bohanec and B. Cestnik (1994), A Schema for Using Multiple Knowledge, *Computational Learning Theory and Natural Learning Systems*, Vol. 2, MIT Press, pp. 157-171.

[22] M. Gams, M. Drobnič and M. Petkovšek (1991), Learning from Examples – A Uniform View. *Int. Journal for Man-machine Studies* **34**, pp. 49-89.

[23] M. Gams, N. Karba and M. Drobnič (1993), Average-Case Improvements when Integrating ML and KA, *Proc. of IJCAI'93 Workshop: Machine Learning and Knowledge Acquisition*, France, pp. 79-95.

[24] M. Gams and V. Križman (1991), The Principle of Multiple Knowledge, *Informatica* **15**, pp. 23-28.

[25] M. S. Gazzaniga (1989), Organization of the Human Brain, *Science* **245**, pp. 947-952.

[26] J. E. Hopcroft and J. D. Ullman (1979), *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley.

[27] D. B. Lenat (1989), When Will Machines Learn?, *Machine Intelligence* **4**, pp. 255-257.

[28] D. B. Lenat (1995), CYC: A Large-Scale Investment in Konwledge Infrastructure, *Communications of the ACM* **38**, pp. 33-38.

[29] D. B. Lenat and R. V. Guha (1990), *Building Large knowledge-Based Systems: Representations and Inference in the Cyc Project*, Addison-Wesley.

[30] N. Littlestone and M.K. Warmuth (1991), The Weighted Majority Algorithm, Technical Report UCSC-CRL-91-28, USA.

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

[32] R. S. Michalski and G. Tecuci (1993), Multistrategy Learning, *IJCAI-93 Tutorial T15*, Chambery, France.

[33] D. Michie (1993), Turing's Test and Conscious Thought. *Artificial Intelligence* **60**, pp. 1-22.

[34] M. Minsky (1987), *The Society of Mind*, New York: Simon and Schuster.

[35] M. Minsky (1991), Society of Mind: A Response to Four Reviews, *Artificial Intelligence* **48**, pp. 371-396.

[36] MTCM (1992), The Machine that Changed the Word, TV series, ACM.

[37] N. J. Nilsson (1991), Logic and Artificial Intelligence, *Artificial Intelligence* **47**, pp. 31-56.

[38] R. Penrose (1989), *The Emperor's New Mind: Concerning computers, minds, and the laws of physics*, Oxford University Press.

[39] R. Penrose (1990), Precis of the Emperor's New Mind: Concerning computers, minds, and the laws of physics, *Behavioral and Brain Sciences*, **13**, pp. 643-705.

[40] R. Penrose (1994), *Shadows of the Mind, A Search for the Missing Science of Consciousness*, Oxford University Press.

[41] J. L. Pollock (1989), *How to Build a Person: A Prolegomenon*, MIT Press.

[42] J. R. Searle (1982), The Chinese Room Revisited, *Behavioral and Brain Sciences* **8**, pp. 345-348.

[43] S. M. Shieber (1994), Lessons From a Restricted Turing Test, *Communications of the ACM* **37**, pp. 70-78.

[44] A. Sloman (1992), The Emperor's Real Mind: Review of the Roger Penrose's The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics, *Artificial Intelligence* **56**, pp. 335-396.

[45] M. J. Stefik and S. W. Smoliar (ed.) (1993), The Commonsense Reviews, *Artificial Intelligence* **61**, pp. 37-181.

[46] A. M. Turing (1936), On Computable Numbers with an Application to the Entscheidungsproblem, *Proc. London Math. Soc.* **2**, pp. 230-265.

[47] A. H. Vera and H. A. Simon (1993), Situated Action: A Symbolic Interpretation, *Cognitive Science* **17**, pp. 7-48.

[48] M. W. Wilkes (1992), Artificial Intelligence as the Year 2000 Approaches, *Communications of the ACM* **35**, pp. 17-20.

[49] T. Winograd (1991), Thinking Machines: Can there be? Are We?, *The Boundaries of Humanity: Humans, Animals, Machines*, Berkeley, University of California press, pp. 198-223, (ed.) J. Sheehan, M. Sosna. Also in this issue.