Biting the apple: the challenge of Artificial Intelligence

Yurij Castelfranchi (castel@sissa.it)
School of Science Communication –
International School for Advanced Studies (SISSA), Trieste, Italy
& LabJor/UNICAMP

Introduction

In a fragment of a love poem, the poetess Sappho of Lesbos (c. 620 B.C.) writes about a woman who was going to be married at not such a young age:

Like a lone red apple,
you stay there, high atop the highest tree…
The apple pickers left you there. Did they miss you?
Oh no, they did not miss you. They couldn't reach that high…

In 1954 Alan Turing, mathematician, war hero, killed himself by biting an apple dipped in poison. Just a few years before, he had thrown to the world an apple of discord, conceiving the possibility of an "electronic child", a programmable machine that could learn, think, be intelligent and even have a conscience. "Can machines think?" – he had written. He thought that yes, machines would think one day. Today, some scientists claim that real knowledge, intelligence and thinking are impossible for machines to reach. Others say that Artificial Intelligence (AI) represents a forbidden fruit: an immoral, and unethical goal. Some say that the apple of knowledge and intelligence is still unripe for computers. Others claim that Artificial Intelligence is just above us, like Sappho's apple: red, ready but too high to reach. The debate is still open. In the meanwhile, machines have learned to learn and to play chess, to demonstrate theorems and to make good medical diagnoses, to interact with us and among each other, and to do many things that were once considered typical of intelligent creatures. In a historical perspective, we briefly analyze here the main steps, the actual state of the art of this ancient challenge and young discipline and the main arguments for and against the possibility of building artificial intelligent beings.

1. Man creates the man: a prehistory of AI and robotics

The challenge (and the fear) to build bodies and minds is not unique to modernity. Artificial life is an old dream of both non-technological and technological cultures. It has been treated as a mythical topic, a magical creation and even a technological and scientific goal.

The word *golem* appears only once in the Bible, as if it was terrible to pronounce. In Psalms 139 we read: "…Thine eyes did see my golem, yet being unperfect. All the days ordained for me were written in your book, before one of them came to be". Golem is used with the meaning of a shapeless mass, an unformed or incomplete substance: such was the clay Adam was made of, before the Almighty gave it life and soul. During the Middle Ages, the golem is transformed into an amazing legend: an artificial being, with neither soul nor sentiments, created by powerful rabbis by means of the cabala. In 1200, German cabalists transmitted the legend of the golem as a clay statue to which two mystics gave life writing on its head the Hebrew word "emet", truth. In the 16^{th} Century, similar legends were told about rabbi Elijah of Chelm, who managed to make a golem using the secret, unpronounceable name of God. Another version tells the story of a real, celebrated rabbi and scientist, Judah Loew ben Bezalel, who lived in Prague and, according to the legend, made a golem to defend the Jewish community from the pogroms organized by the Christians. But the monster turned out to be a terrible problem and had to be eliminated by deleting the first letter – an aleph – on his forehead. "Emet", truth, turned to "met", "death": only God, says the moral lesson, can create life. The legend spread all over Europe and later was made into novels, poems and movies.(1)

The quest for automata and artificial humanoids was not only rooted in Jewish culture. Classical Greek poets sung the myth of Pygmalion, who carved the statue of a beautiful woman, so perfect that he fell in love with his own creation. The Gods, touched by the prayers of the powerful artist and craftsman, gave life to the artifact. In the *Iliad* Homer tells us about the wonderful house of Hephaestus (or Vulcan, god of fire, craftsman and blacksmith), where tables with three legs could walk and where live golden statues worked as waitresses to serve wine. Later, during the Hellenistic Era, mythology was no longer the only place in culture in which automata could be present. Several authors emphasized the importance of what we call today "pure research" and technology in the Hellenistic culture, and how during this epoch the seeds were planted for what we consider "modern" science.(2) In the First Century B.C., Hero of Alexandria built the first steam engines recorded in history and used them to achieve amazing automata, which was later re-invented during the Renaissance.

Automated puppets were not the only simple dream toward artificial beings. The alchemists dared to imagine the creation of life and humanoid beings. The possibility of creating a *homunculus* dates at least from the Gnostic groups, who, in the Third Century A.D., merged Christian religion with oriental influences and magical knowledge coming from Greek and Roman cultures. Centuries later, Theophrast Bombast von Hohenheim, aka Paracelsus (1493-1541), a celebrated alchemist and physician,[3] even gave a recipe to create *homunculi* based on human sperm, forty days of fermentation and the heath of a horse's womb. Wolfgang Goethe, who loved science and magic, physics and alchemy, wrote about colors theory and sorcerer's apprentice, transformed in his *Faust* the homunculus into a gentle, little humanoid.

During the Middle Age, myths about automata and artificial beings took on new life and spread throughout Europe with new connotations. At the end of the 13th Century in England, the weird legend was spread of a brass talking head, built by Roger Bacon (1214-1294), a philosopher who prophesized the coming of an era dominated by machines. Actually, thanks also to the ancient mythical connotations that link knowledge with power (and magic)[4], many famous scholars and savants were credited by legend with the ability to build artificial talking heads. Roberto Grossatesta (1168-1253), a mathematician and philosopher, had one of those heads, as well the "magician" pope, Silvester II (alias mathematician Gerbert of Aurillac). Even St. Albertus Magnus (1206-1280), also known as Albert the Great, was credited with having built a robot. The saint, who was also a Bavarian philosopher, theologian and bishop as well as geographer, astronomer and chemist, was famous among his contemporaries as the "universal doctor", or "the wonder of knowledge". A legend tells that he was able to organize a banquet during which the guests were served by metallic waiters. Later, similar tales were told about several other savants. The great mathematician Johan Müller (1436-1476), known as Regiomontanus, was credited as the inventor of an artificial eagle able to fly into the skies of Nuremberg and welcome the emperor Maximilian. According to a similar story, Leonardo da Vinci constructed a mechanical lion in Milan in 1499 to honor Louis XII of France: when the king entered the city, the lion moved forward, opened its chest with a claw, and showed a lily, the symbol of France. Even Renée Descartes was a victim of such stories: he was credited as the builder of a robot that he called *Ma fille Francine* ("my daughter Francine").

While these legends were spreading throughout Europe, people began to believe that it was actually possible to build artificial bodies. In 1540, Giannello Della Torre, a craftsman from Cremona, built for Emperor Carl V a mechanical young lady who was able to walk and play music by picking the strings of a real lute. Isaac and Salomon de Caus, French garden designers, built several "robot" fountains at the beginning of the 17th Century. The most famous of them had a bird that could sing, thanks to a hydraulic mechanism, and escape when a robot-owl appeared. In Japan in 1730 a technical manual about automata technology known as the *Karakuri kimon kaganigusa* was printed. The most famous automata of the epoch are those by Pierre Jacquet-Droz and Jacques de Vacaunson. In 1774 Jacquet-Droz built an automatic writer which still lives in a museum in

Neufchatel (Switzerland). It puts its pen in the inkstand to write a real letter of two lines and, what is more amazing, the letter is programmable by the user. Vacaunson, on the other hand, was famous because of a copper duck that unfortunately disappeared. According to several witnesses, it was able to swim, drink, quack and even swallow grains, which were chemically digested thanks to substances contained in its cooper intestine. The duck, which even defecated at the end of the show, traveled in tours all over Europe and was admired by Voltaire and Goethe. It consisted of eight hundred different pieces and Vacaunson (who also built some automatic musicians) opened the robot after each show to demonstrate the mechanisms.

All these artificial beings, of course, had nothing to do with mind or intelligence: even if some of them played some role in the prehistory of robotic mechanisms, they were merely sophisticated puppets, made with the intention to amaze. However, gears and wheels were soon used not only to imitate parts of living bodies, but also to try to achieve another ancient dream: to mechanize mathematical calculations and logical reasoning. In 1642 Blaise Pascal (1623-1662), a 19 year-old boy who was to become one of the greatest scientists and philosophers of the century, built a calculator-box as a present for his father, a French tax collector. The box was a small masterpiece: it could calculate additions and subtractions of up to 8-digit numbers.(5) The Pascal calculator was improved by another great master of science. Gottfried Wilhelm Leibniz (1646-1716), philosopher, mathematician, and logician (and father, together with Isaac Newton, of infinitesimal calculus), modified the Pascal machine to allow it to perform multiplications and divisions. Even if his machine was only a calculator, Leibniz understood that, in the future, machines would also be able to perform logical operations. Leibniz also worked as a diplomat and his dream was to develop a formal "perfect" language, which one day he wrote, philosophical and even political discussions would be solved simply by calculation. Some decades later, an idea by de Vaucanson for an automatic loom was improved by Joseph-Marie Jacquard, who invented punch cards and built a programmable loom. The world was ready for the idea of machines, which were able to learn. Something very similar to what we now call a computer was going to be invented by an eccentric mathematician and the daughter of a rebel poet.

2. A mechanical computer

They were definitely a strange couple. They had met on a November evening in 1834 at a dinner organized by Mary Sommerville, mathematician and translator of Laplace's works into English. Ada was the daughter of the celebrated poet Lord Byron, but she had never met her father. Charles, an eclectic, eccentric scientist and a member of the Royal Society, went to the dinner to present his new invention: a calculating machine. Not a simple one, but a powerful, “analytical engine”, able to calculate logarithms and square roots, planetary orbits and routes for ships. Nobody paid much attention to Charles Babbage’s presentation. Nobody that is, but the young Ada Byron. She would marry another man (the Count of Lovelace) but also be Babbage’s inseparable scientific partner. Babbage was to become one of the inventors of the computer. Ada Byron was to become the first programmer in history.

Actually Babbage had already projected a “difference engine” several years before, which was able to perform several mathematical calculations. He never managed to transform the project into reality, but nevertheless decided to be even more ambitious with the analytical machine, the idea of which was very similar to that of a modern computer. It should have five main components:

- a “store”, or memory (about 50 Kb in size in present language, but virtually infinite because any amount of information could be stored, read and written on punch cards);
- a “mill”, the place in which data were manipulated (the equivalent of our processor, or CPU);
- a “control”, written on a punch card, with the list of operations to perform on data - it is what we refer to today as the program; and
- “input” and “output” devices (Babbage even thought of mechanisms to print out the data).

In the words of Babbage himself, Ada understood “all these things” better than he did and was much more capable of explaining them. In her writings, Ada merged logic and mathematics and understood the importance of using the punch cards as a list of commands. She said that, like a loom that wove pictures, leaves and flowers, the programmable logic machine could “weave algebraic schemes”. She wrote some simple programs for the analytical engine, and conceived the very base of what makes a computer different from an abacus or a mechanic calculator: the conditioned instructions (IF… THEN). The engine could learn and modify its behavior according to external data and situations.

Charles and Lady Ada never managed to build a complete, functioning version of their engine. Today, we know that it was possible. After Babbage’s attempt, the prehistory of AI and computers becomes history and it is relatively well known. The mathematician Alan Turing was a war hero: he and his team managed to crack the famous Nazi Enigma machine used to code messages, a fact which played a crucial role in the resolution of the Second World War. A few years later, according to UK law of that time, he was a criminal because he was gay. Condemned to forced hormone “therapy” (which changed his voice and caused his breasts to grow), he committed suicide in 1954 by biting an apple dipped in cyanide.

3. "Can machines think?"

The story of Alan Turing is too complex and rich to be told here. We summarize only part of his intellectual route. In 1936 the young Turing, a student at King's College in Cambridge, managed to solve one of the big, difficult problems posed by David Hilbert in 1928 as a challenge to the mathematics community: the *Entscheidungsproblem*, or "decision problem". The problem was relatively simple to enunciate, but very difficult to solve: is it always possible to "decide" if a mathematical enunciation is true or false? Turing managed to solve the problem(6) (the answer is no, it is not always possible to establish the truth or falsity of a mathematical theorem), by imagining the existence of a machine,(7) a "logical computing machine", known today as the "Turing universal machine". We will not enter into details of the Turing demonstration. What is important to note here is that, with the idea of a logical universal machine, Turing also showed the logical existence (soon to become concrete existence) of machines able to perform any kind of activity describable in terms of an algorithm. He was given the opportunity to study for his PhD at Princeton University, where some of the greatest mathematicians and logicians were working: John von Neumann, Godfrey Hardy, Richard Courant, Albert Einstein, and Kurt Gödel. But Turing didn't stay long. In 1938, when the war was about to begin, Turing decided to go back to his country. He played a crucial but secret role in the Government Code and Cypher School. Adapting and improving some electromechanical machines developed by Polish mathematicians, he began to decipher German codes and to build a huge calculator known as the Colossus. In March 1943, the Colossus managed to decipher the messages from the Nazi naval headquarters. In 1946 Alan Turing was awarded the Order of the British Empire medal for his achievements, but he could never explain why: his activities were considered top secret until twenty years after his death. In the same years, on the other side of the Atlantic, John von Neumann, working on the Manhattan Project to build atomic bombs, contributed to the creation of some of the first electronic programmable computers, Eniac and Edvac,(8) and invented the first forms of Artificial Life: cellular automata. Von Neumann strongly felt that parallels existed between computers and the brain. In his articles, he began to use the term "electronic brains" to indicate computers and claimed that the machines could be used as a tool to study biological life and the mind. Turing was even more ambitious. After the war, he began to tell his collaborators that they were going "to build a brain". In 1947 he declared that computers were still used only as slaves, able only to obey fixed lists of commands, but that this was not the only possibility: computers could learn, auto-modify their programs, do things which nobody programmed them to do. If we could give a computer not only a "brain" but also the possibility to interact with the environment, manipulate things and learn from the real world, then one day, Turing said, we could eventually have a "child machine".

In 1950, in one of his most famous works,[9] Turing wrote, "I propose to consider the question 'Can machines think?'. This should begin with definitions of the meaning of the terms 'machine' and 'think'. […] Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. The new form of the problem can be described in terms of a game which we call the 'imitation game'". The game, later known as the "Turing Test" (although Turing actually didn't think of it as a final test for machine intelligence), is played with three people: a man (A), a woman (B) and an interrogator. The interrogator stays in a room apart from the other two. The object of the game is for the interrogator to determine which of the other two is the man and which is the woman, communicating by means of a "teleprinter" (today it could be any kind of electronic chat device). Turing writes: "What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original: 'Can machines think?'". Turing wrote: "the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking. One could then describe these feelings to the world but, of course, no one would be justified in taking any notice. Likewise according to this view the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe 'A thinks but B does not' whilst B believes 'B thinks but A does not'. Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks". So, Turing concluded, if a machine can perform the same kind of mental processes that a human performs, in such a way that we don't manage to recognize if we are talking to a human or a machine, we should admit that it thinks. "I believe that in about fifty years time" – he wrote – "it will be possible to programme computers […] to make them play the imitation game so well that an average interrogator will not have more than 70-per cent chance of making the right identification after five minutes of questioning. […] 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".

He was too optimistic. Fifty years passed, but machines, according to general opinion, don't think. However, the bases for AI, and the apple of discord, were thrown. Today machines do many things that were considered unique to intelligent, thinking beings.

4. AI is born

The name of the new discipline, "Artificial Intelligence", was to be born only after Alan Turing's death. In 1956, at Dartmouth College, Hanover, USA, mathematician John McCarthy organized a workshop that he called "Summer Research Project on Artificial Intelligence". Many of the brilliant minds who would become the "fathers" of AI were invited to the event: Claude Shannon, founder of the information theory and his pupil, Marvin Minsky, soon to be a star of the new field; Walter Ashby; Alex Bernstein; John Holland; Donald McCulloch and many others. In their *manifesto*, the participants declared to be united by the belief that machines would, one day, perform tasks traditionally considered intelligent. Since then, several definitions have been proposed for Artificial Intelligence.[(10)] For some, AI is the area in computer science that tries to build "computers that behave like humans". For others, AI is simply "what computers don't do yet". The most cautious define it as "the study of how make computers do things that, if done by humans, would be considered intelligent tasks ".

At Dartmouth College, three scientists brought a precious program as a present to the newborn AI. Cliff Shaw, Allen Newell and Herbert Simon (who would be a Nobel Economics Prize winner in 1978) came not only with ideas, diagrams, and projects about how to build thinking machines, but also with a real, functioning intelligent program, able to demonstrate and discover mathematical theorems. The *Logic Theorist* was programmed to demonstrate theorems based on the framework of the *Principia Mathematica*, the celebrated work by Alfred North Whitehead and Bertrand Russel. One year later, Newell and Simon produced a more ambitious program, the *General Problem Solver*, which was able to solve several kinds of mathematical problems and to demonstrate general theorems.

4.1. AI today: the state of the art

Two main approaches in Artificial Intelligence compete today in the quest for thinking machines. The "classical AI" vision involves mainly "top-down" strategies: algorithms and programs to manipulate symbols and concepts, to plan actions and simulate different aspects of perception and reasoning. The "bottom-up" approach, supported by Artificial Life, genetic programming, and neural networks, involves learning by trial and error and reactive processes that are not necessarily planned by symbolic rules or reasoning. Some scientists belonging to the latter area do not define this approach as a part of AI but as an opposite way toward intelligent artificial beings. We briefly resume here some of the main fields of application for AI and describe, in a very simplified way, the approaches.

4.1.1. Playing games

Since the very beginning of AI, Alan Turing foretold that an important skill machines should develop was playing games. Turing himself wrote a simple first-draft program for playing chess and was very upset that it didn't win playing against him. In 1950 Claude Shannon, using the minimax theorem, published the first important work on chess-playing computers.(11) Seven years later, Alex Bernstein produced the first complete software for chess. In 1967, the program *MacHackVI*, written at MIT by Joseph Greenblatt, beat a human player in a championship. With increasing memory and speed, personal computer programs were playing better and better. Finally, in May 1997, *Deep Blue*, an IBM computer, won against the human world champion, Garry Kasparov. Today, computer programs play Otello better than any human and they play top level chess and checkers. But they are still mediocre or very bad players in games (like Go, bridge or poker) in which intuition cannot be replaced by brute force.

4.1.2. Expert systems and learning systems

Since the times of the first chess-playing programs and of *Logic Theorist* and *General Problem Solver*, Artificial Intelligence has passed through several phases of success and euphoria as well as many "winters" of disillusion and crisis. The '60s and '70s were a hot "summer" for AI. Along with Bruce Buchanan and Georgia Sutherland, Edward Feigenbaum and Joshua Lederberg (Nobel Medicine Prize winner) created *Heuristic Dendral*. It was able to reconstruct the chemical structure of a substance starting from mass spectrometry data, just like a human expert. A few months later, Joel Moses invented *Macsyma*, an expert in solving indefinite integral calculations. Expert systems, which belong to the family of the knowledge-based system, are composed of two main parts: a "knowledge base" (which contains the huge set of rules that a human expert uses, consciously or not, to solve a problem) and an "inferential engine" (which includes the logical rules used to deduce a new fact from known information and assumptions). Expert systems were successful, both scientifically and economically, in the '70s and were marketed as the example of real intelligent artificial systems. Actually, their performance can be quite dramatic. In 1975, the *Dendral* discovered a new rule, unknown to human scientists, to identify some kinds of organic molecules. In 1979, *Mycin*, a program which is an expert in the diagnosis of blood infectious diseases, and *Internist*, an expert in internal medicine, were shown to be capable of making diagnoses at the level of the best physicians in the field. Three years later, *Prospector*, a geological expert system, discovered a rich deposit of molybdenum. In 1986, hardware and software business connected with AI reached 425 million dollars, largely due to expert systems. Later, the idea evolved. Expert systems were created with the ability of learning and creating their own rules in knowledge base. *Id3*, a famous program of this kind written by Ross Quinlan, was used to control nuclear reactors. *Assistant* was able to diagnose cancer better than many physicians while *Golem*, by Muggleton and Feng, was used to calculate the tridimensional form of proteins.

Today, expert systems are still used. Factories use them to control part of the machine operations. Expert systems exist today for the repairing of computers and cars, military analysis of battle field conditions, surgery, oil perforation, financial planning, weather forecasting, newspaper make up, tax declarations, etc. *Tacair-Soar*, a sophisticated military expert system with over 5,000 rules (and that also integrate other AI approaches), is able to pilot war planes in an intelligent way, creating evasion tactics and using missiles and bombs.(12) But are expert systems really "intelligent"? No doubt, they share some characteristics with intelligent beings: they are able to solve complex problems, and they do this by fragmenting problems into smaller ones; they "understand" when we give them incomplete or contradictory information; their knowledge is not fixed and static: they can learn. On the other hand, expert systems are definitely stupid entities. They completely lack some fundamental features of intelligence: they are very rigid, completely unable to cope with something outside their area of "expertise"; they have no creativity, they cannot imagine new strategies to solve new problems; they have no idea of the external world: they have no "common sense". A good expert system used to evaluate loans for a car dealer gave a fantastic loan to a 20-year-old boy who declared to have 10 years of work experience: the software did not have any idea of what being 10 or 20 years old meant and of what labor laws were.

In 1982, Douglas Lenat tried a different approach to learning systems. His *Automatic Mathematician* was able to "discover" principles and properties in mathematics. In one night, the program, which just knew the definition for "class", "union", "intersection" and some logical rules to organize concepts, "discovered" natural numbers, addition, multiplication and prime numbers. It even invented the opposite concepts of prime numbers, unknown to Lenat himself. In 1984 Lenat created *Cyc*, a huge intelligent program that has some degree of "common sense" and knowledge about the real world. With over one million rules in its database, *Cyc* "knows" that when someone is dead he usually stays dead, that things exist which can be touched and others which cannot, and so on. Lenat claims that in some twenty years *Cyc* will know tens of millions rules and will begin to have real consciousness and intelligence.

4.1.3. Natural language

The ability to have a conversation in our usual daily language is the crucial part of Turing's idea for the imitation game. Since the beginning of AI, lots of funds and energy have been invested to try to build efficient automatic translators and programs able to understand and communicate with us in natural language. One of the most famous first attempts to write a program with good communication skills is *Eliza*, by Joseph Weizenbaum. Written between 1964 and 1966, it simulates a rogerian psychoanalyst. It was soon world famous and some individuals actually wanted to be *Eliza*'s patients. A typical conversation with the electronic doctor was of this kind:[13]

Human: Men are all alike.

Eliza: In what way?

H: They're always bugging us about something or other.

E: Can you think of a specific example.

H: Well, my boyfriend made me come here.

E: Your boyfriend made you come here.

H: He says that I am depressed all the time.

E: I am sorry to hear you are depressed.

H: It's true that I am unhappy.

E: Do you think coming here will help you not to be unhappy?

H: I need help that seems certain.

E: What would it mean if you got some help?

H: Perhaps I could learn to get along with my mother.

Later, a "paranoid" software was also created: *Parry*, by Kenneth Colby, was able to simulate a simple conversation and some human doctors actually believed they were talking with a real patient:

Doctor: Why are you in the hospital?

Parry: I shouldn't be here.

D: Who brought you here?

P: The police.

D: What trouble did you have with the police?

P: Cops don’t do their job.

D: What do they do instead?

P: Cops arrest the wrong people. […]

P: You don't believe me?

D: I am doing my best.

--*Parry* remains silent--

However, these programs could not really understand what was said. They just picked up discourse fragments and replicated them, giving the impression of real, empathetic listening. In 1991 millionaire Hugh Loebner created a prize, to be awarded every year, for the best programs able to confront Turing’s imitation game. He offers one hundred thousand dollars for the program that is able to pass the “Turing Test”. No program has done so yet. But every year a smaller prize is offered for the “most human-like program” (in addition to a joke prize for the human with the worst human-like performance, according to a jury). However, professional software for natural language and translation is quite sophisticated today and can handle very good conversation, if it is about very specific topics (for example, hotel reservations, molecular biology, etc.), and also solve complex problems posed by the user. One of the very first attempts of such software was SHRDLU, which lived in a simple toy world made by colored cubes, pyramids and spheres. This is a real conversation with SHRDLU:

Human: Pick up a red block.

Shrdlu: OK!

H: Find a block, which is taller than the one you are holding and put it into the box.

S: By "it" I assume that you mean the block, which is taller than the one I am holding. OK.

H: What does the box contain?

S: The blue pyramid and the blue box […],

H: Will you please stack up both of the red blocks and either a green cube or a pyramid?

S: OK […].

H: Had you touched any pyramid before you put the green one on the little cube?

S: Yes, the green one.

H: When did you pick it up?

S: While I was stacking up the red cube, a large red block and a large green cube.

Today, several systems of this kind can solve real life problems and interact with us almost as a person would. *Alfresco* (Automatic Language-Fresco Interactive System), developed at the Istituto per la Ricerca Scientifica e Tecnologica, in Trento (Italy), can talk with us about 14th Century Italian art. It can understand very complex questions and answer producing complex texts, showing pictures and movies:

Human: Show me a painting of Giotto in the Cappella degli Scrovegni.
Alfresco: [the system shows *Fuga in Egitto* on the touch screen]
H: Is there a painter who was influenced by the master?
A: Yes, for example Ambrogio Lorenzetti.
H: Who is this baby? (the user touches the image of Infant Jesus on the touch screen)
A: Jesus.
H: Where was Lorenzetti born?
A: In Siena.
H: Show it to me.
A: [the systems shows a brief movie about Siena]

4.1.4. Robotics and neural networks

Cybernetics, first developed by Norbert Wiener (1894-1964), and neural networks (invented by Warren McCulloch and Walter Pitts in 1943 and later developed by Donald Hebbs and many others) played a crucial role in expanding the horizons of AI and providing ideas for new approaches toward the building of artificial intelligent beings. In 1980, about 20,000 robots worked in factories all over the world. Today there are at least 200,000, 60% of which "live" in Japan. In the '80s a new generation of intelligent robots was developed, supporting a new approach for AI. Instead of complicated software planned to make computers understand rules and properties of the world and deduce behaviors (top-down approach), several scholars developed a bottom-up approach, based not on complicated algorithms, but on reactive systems that manage to learn and modify their behavior. Rodney Brooks of MIT has been a pioneer of this new approach. Claiming that "elephants don't play chess", he asserts that we can build "intelligence without reasoning". His robots, like *Genghis* and *Attila*, are electronic arthropods that do not manipulate symbols, do not use logic, and do not have any intelligent software for planning. They are based on reactive neural networks and learn how to move and interact by trial and error. The small robots sent by NASA to explore the surface of Mars are a good example of the success of this kind of approach. Neural networks are also used to produce speech and pattern recognition systems and many other learning devices. In 1995, *Ralph*, intelligent software for artificial vision developed at Carnegie Mellon University, managed to drive a car from Pittsburgh to San Diego with no human help during 98% of the journey.

Brooks continues to challenge the view that this bottom-up approach cannot produce real artificial minds. He is currently trying to build electronic children based on this architecture. *Cog* (cognitive robot) is trying to understand the world through exploration with its single arm and hand. *Kismet* also has facial expressions and can express curiosity, sadness and boredom, based on the stimuli it receives. In the meanwhile, robots have been built that can walk, play the flute and the piano, play soccer (at the RoboCup world championships, for example) or karate, work as nurses or butlers. Robot dogs and cats can play with our children.

Another sub-symbolic approach to artificial intelligence is that of genetic programming, invented by John Holland in the '80s. The idea is based on Darwinian evolution: simple algorithms are generated, they reproduce themselves automatically with some small "mutations", and only the ones best suited to cope with a defined goal survive. The others are deleted. More generally, a big research line opposed to classical AI and based on bottom-up approaches was developed starting from Von Neumann's cellular automata. Today it is called Artificial Life and it has been used, for example, to produce virtual characters for video games and movies.

4.1.5. Distributed Artificial Intelligence

In recent years, classical Artificial Intelligence also developed a powerful new approach. Instead of thinking of an intelligent machine as a huge computer (like Hal 9000 in Stanley Kubrick's *2001: A Space Odyssey*), people began to understand that intelligent behavior could be an emergent property of a complex network based on simple agents. At the end of the '70s, Marvin Minsky was already one of the pioneers of this new approach.(14) Today, Distributed Artificial Intelligence and intelligent agents are used successfully everywhere: "electronic virtual ants" are used to solve problems in telephone networks. Agents are autonomous: they can interact with other programs, negotiate for us, and ask for details if they lack information. Agents are social, too: their behavior is based on rules, goals, and interests. Barbara Grosz developed intelligent agents that manage electronic commerce, distance learning, and group collaboration management.(15). Intelligent agents today can use web tools (like mail, ftp, gopher, telnet, archie, etc.) and do intelligent data mining for us, buy on the net the best car according to our budget and style, find special prizes and make travel reservations, and so on. Maja Mataric mixed Brooks' bottom-up approach with the intelligent agent idea to develop social robots that work in cooperation to perform difficult tasks.

4.1.6. Toward affective computers

In recent years, several groups around the world have worked towards reaching one of AIs most difficult goals: to build intelligent systems able to recognize, simulate or even feel real feelings. Rosalind Picard of MIT thinks that it is possible to build “affective computers”. Her group is connecting computers to instruments similar to lie detectors and programming them in order to recognize the feelings of the user. Other groups have managed to produce software able to respond to users’ feelings, expressing emotion themselves by means of virtual faces. The *Affective Reasoning Project* at DePaul University built graphical agents that can listen, show faces, and answer with an emotional voice. Manfred Clynes, a neuroscientist from California, has produced programs able to execute music scores with emotional characteristics chosen by the user, while Michael Casey of MIT has written software that can play together with a musician, “feeling” the emotional tone of the music. Several other groups, especially in Japan and the USA, have produced software pets and virtual characters with emotional characteristics for movies, cartoons, and video games: the *Woggles* (by the Oz group, at Carnegie Mellon), the *Norns* (by CyberLife) and the *Petz* are examples of applications of this approach. MIT’s Alex Pentland is producing “wearable computers” that can perceive our feelings (anger, fear, stress, and joy) by means of sensors and eventually suggest solutions. Several affective systems have been made to work as funny teachers or educators. *Herman the Bug*, created at North Carolina State University, is a virtual insect that gives funny lessons in biology. *Adele* (Agent for Distance Learning Environments), born at University of Southern Carolina, assists and monitors students in distance learning courses. Of course, all these applications suggest reflections about the real necessity for this type of application and pose the question of how social control and social life can be deeply modified by technology. And, of course, none of these machines can be said to really feel feelings. But many researchers in the area think that this is not an impossible task.

5. The apple of discord: is AI impossible?

In 1957 Allen Newell and Herbert Simon claimed that in ten years machines would be able to write beautiful symphonies, play chess better than any human being, and explain psychology. Some years later, philosopher Herbert Dreyfus was sure that the whole dream of AI was nothing more than crackpot science and that no computer would ever play chess well. Both of them were wrong. In 1965 Simon had to admit that the ultimate goal of AI could not be reached quickly, while Dreyfus was defeated, in 1967, by a quite simple computer program named *MacHack*. In his most celebrated book – *What computers can't do*(16) – the philosopher rebutted that not much intelligence was needed to win against a mediocre opponent such as himself and that the possibility that a computer program could play really well was only "scientific mythology".

In 1970, Marvin Minsky of MIT declared that in three to eight years he expected to see machines with an IQ comparable to that of an average human being: computers able to read Shakespeare, wash a car, tell a joke, become upset, etc. Soon after, he explained, these machines could begin to learn and educate themselves, quickly becoming more intelligent than any biological thing on earth. Dreyfus replied that this could happen neither in eight years nor in eight centuries: it was simply impossible. The AI challenge, the philosopher said, was like children climbing a tree in an attempt to reach the moon. Again, both were wrong. Supporters of Artificial Intelligence were too quick in promoting the discipline and promising huge leaps of progress that never happened. On the other hand, opponents of AI were often too naive in trying to demonstrate "scientifically" the *a priori* impossibility of any intelligent behavior in machines. They negated every step of progress, changing their definition of what mind, consciousness and "real" intelligence was every time that a machine could perform a job previously thought to be typical of humans: if a machine could perform "X", then "real" intelligence was only "Y", that was "impossible" for a computer to achieve. If "Y" was finally achieved, then certainly a machine could never really perform "Z" well, and so on. Today, computer and robots, like Alan Turing foretold, can see, touch, hear, learn, communicate with us and among themselves, solve problems, explore the world, make predictions, etc. However, no software can satisfy the euphoric expectations of AI supporters: no machine can have a real, complete conversation with us, no robot can interact with the environment with the same talent as that of a young dog. There is no computer that can really have fun with a joke or become sad. No computer knows it exists yet. So, who's right? Computers don't think *yet* or will they *never* think? Is our brain a computer and a computer can carry out all functions of a human brain? And if so, is all of what makes our "I" in our physical brain, or does our mind contain some ingredient that a machine will never have? Supporters of the "strong AI" claim that at the base of our consciousness there is the capacity of our brain to manipulate symbols and transform information. Since computers can do the same, nothing forbids, at least in principle, the possibility that a machine can have thoughts, feelings, and consciousness. On the other hand, advocates of the "weak" version of AI think that software cannot have all of and exactly the same functions as an intelligent mind: maybe some other kind of machine will in the future, but computers today are just useful tools to investigate and perform functions similar to some of those which are typical of our intelligence. Other scholars oppose the idea of AI in any form: they think that the very idea of intelligent machines is mathematically or physically impossible, philosophically incoherent, or morally unacceptable. And they have tried to demonstrate this.

3.1. Is AI "mathematically impossible"?

In 1930, Kurt Gödel (1906-1978), Austrian logician and mathematician, demonstrated his two incompleteness theorems, making history as one of the greatest logicians by overcoming one of the most difficult challenges posed by David Hilbert in 1900 and giving a final blow to the project of "mathematizing" the world. Thirty years later, his theorems would be used in an attempt to demonstrate the impossibility of AI.

Gödel's first theorem, stated in a very simplified form, claims that every formal system sufficiently elaborated (like, for example, our arithmetic) is either inconsistent (meaning that it allows to demonstrate false theorems) or incomplete (that is, it is unable to decide the falseness or truth of some theorems). Actually, Gödel also spoke about AI. He believed it was an impossible task to achieve: he thought that our brain is indeed a machine, but that our mind does not only consist of a brain. Something transcendent was linked to our physical brain, something that a machine could not have and that made us able to, for example, grasp the idea of infinity. But he never claimed to have demonstrated mathematically the impossibility of AI. Other scientists later used Gödel's first incompleteness theorem to try to do this. One of the first was J. Lucas: his paper, *Minds, Machines and Gödel*, was improved by Paul Benacerraf in 1967 and finally re-elaborated in recent years by Roger Penrose.(17) Their point is complicated to demonstrate mathematically, but the idea is quite simple. Let us suppose we can build an algorithm as complicated as we want to imagine it, able to simulate perfectly well all aspects of human thought. However complex and sophisticated, this system will consist necessarily of some formal system (that is, a system based on symbols and rules to manipulate symbols). If this system is not inconsistent (i.e., if it doesn't claim absurd and false things), then we humans, since Gödel's theorem assures it, can find some "sentence" or statement that this system is completely unable to "decide" or demonstrate as true or false. In Gödel's terms we can write, in logical symbols, the equivalent of this claim: "This sentence is not demonstrable". If the sentence is true, then the system is incomplete because it is not able to demonstrate something that is true. If the sentence is false (and can thus be demonstrated), then the system is inconsistent, demonstrating something false. If our intelligent machine is able to simulate the rationality of a human being, then it must be consistent, and the above sentence is true. So, simply by logical insight, a human mind can see that the claim is true, while the machine, which can only manipulate symbols, cannot see it as true nor demonstrate it as false. Therefore, any complex machine will always be different from and weaker than the human mind. "There is some elusive and ineffable quality to human intelligence", explains Lucas: "however complicated a machine we construct, it will, if it is a machine, correspond to a formal system, which in turn will be liable to the Gödel procedure for finding a formula unprovable-in-that-system. This formula the machine will be unable to produce as true, although a mind can see that it is true. And so the machine will not be an adequate model of the mind". In recent years, Roger Penrose, mathematician and physicist, added to these arguments yet another one: the brain is not only regulated by chemistry and usual quantum physics, but also, in his opinion, by the not yet understood laws of quantum gravity. No electronic machine could use this type of physics, which could be the base for the mind. Supporters of AI, of course, as well as many philosophers of mind, didn't remain silent. Logicians such as Hilary Putnam, George Boolos, Martin Davis, Solomon Feferman; cognitive scientists such as Douglas Hofstadter and philosophers such as Daniel Dennet found several flaws in the arguments of Penrose and Lucas.(18) And they stressed them in a very virulent way. According to Putnam, Penrose's books are "a sad episode of our present intellectual life". The main objection to this "Gödel's" argument is that it is based on the assumption that a machine imitating our mind should be consistent. But how can we know that a formal system able to imitate a conscious mind would be consistent? Perhaps our brain builds up a formal system that is so complicated and complex that it admits the existence of several logical absurdities. Moreover, even if we imagine that we humans are logically coherent, maybe we could be incomplete, suffering the same limitations as any other machine: maybe there is a terribly complicated formal claim that is true, but that we could never decide as true or not. Several other more complex mathematical and logical arguments were presented against Penrose. But the main point is simple: Gödel's theorem does not demonstrate that our reasoning and mind are not representable on a computer. It just shows that if they are, then they must obey the same limitations as any other formal system.(19) Penrose answered that indeed we cannot exclude that our mind is incoherent or otherwise unable to demonstrate true theorems, but according to him, this is a very remote possibility: we should already have found such bugs in the mind, if they exist.

3.2. In the "Chinese room": is strong AI philosophically impossible?

If the mathematical attack to AI doesn't seem to have had a big impact on the AI community, another philosopher, the American John Searle, proposed an anti-AI paradox, a thought experiment that deserved a deeper debate. Searle, unlike Lucas and Penrose, is not against the idea of AI a priori. What Searle definitely doesn't believe is the strong AI proposal: the idea that a conscious, intelligent mind can be implemented on a computer by means of an algorithm. "Can a machine think?" – Searle asks. He answers that only machines can think: for example, those very special machines we call brains. But he believes that these machines cannot be computer programs. Software, according to the philosopher, could never have a real understanding of the world. To demonstrate this, Searle invents a mental experiment.[20] Let us suppose that we made a computer that manages to pass the Turing test, behaving exactly like a Chinese person. The computer takes Chinese symbols as input, consults its huge database, uses a terribly complex intelligent algorithm and then produces a Chinese output so perfect that a Chinese speaker is convinced that he or she is talking to another Chinese person. The conclusion that proponents of strong AI would like to draw is that the computer understands Chinese, just as the person does. But Searle asks us to suppose that he is sitting inside the computer. In other words, he is in a small room in which he receives Chinese symbols, executes manually the algorithm, uses the huge data-base and returns the Chinese symbols that are indicated by the program. Searle writes: "I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker […], but I still understand nothing. […] The computer has nothing more than I have in the case where I understand nothing." Contrary to the strong AI point of view, then, no matter how intelligently a computer behaves and no matter what kind of software it is running, the symbols it processes are meaningless to it. The internal states and processes of the machine, says the philosopher, are purely "syntactic". They lack semantics (i.e., meaning). The computer, says Searle, cannot have "really intentional mental states".

Dozens of replies were given to Searle's arguments.[21] The most popular says that though Searle himself doesn't understand Chinese, it is perfectly correct to say that Searle plus the database and rules inside the room understand Chinese. The entire system, claim strong AI supporters, understands Chinese, exactly in the same way as no single neuron or small part in our brain understands anything or knows it exists, but our mind does. No person, actually, could memorize or use such a big algorithm: millions of people would be necessary to perform all the operations and, even if none of them really understood Chinese, the intelligent Chinese person would "emerge" from this collectivity, like the mind emerges from brain activity. Searle rebutted that even if this answer is possible in principle, it represents "a form of intellectual pathology" and that only someone who is a "prisoner of an ideology" could find some plausibility in it. The debate is still on.

Even today no agreement exists on how to demonstrate whether strong AI is possible or impossible in principle. In the meanwhile, several people debate if AI is a correct enterprise from the moral point of view: a bioethics of the artifacts is born.

3.3. Is AI an immoral experiment?

The ethical problem with AI, robotics and computer technology is definitely an extremely complex one, involving dramatic topics such as social control, democracy and empowerment, "human rights" for thinking machines, worker's rights, the social impact of technology, and social representation and image of science and technology. It is a theme, which deserves deeper reflection and more space than that available here. We summarize only some of the main points here.(22)

Already in 1948, Norbert Wiener, in *Cybernetics: or control and communication in the animal and the machine*, wrote the following: "It has long been clear to me that the modern ultra-rapid computing machine was in principle an ideal central nervous system to an apparatus for automatic control. […] We are already in a position to construct artificial machines of almost any degree of elaborateness of performance. […] We [are] here in the presence of another social potentiality of unheard-of importance for good and for evil." This classical, quite naive position about neutrality of pure science versus the "good" or "evil" application of technology became much more complicated later. During the late 1960s, Joseph Weizenbaum, father of *Eliza*, was shocked at the reactions people had to his simple computer program: some practicing psychiatrists saw it as evidence that computers would soon be performing automated psychotherapy. Even computer scholars at MIT became emotionally involved with the computer. Weizenbaum was extremely concerned that strong AI could reinforce the tendency to see humans as mere machines.

Today, some supporters of AI have come to the opposite conclusion: even if AI can be dangerous and very questionable from an ethical point of view, they say it is a legitimate goal for humankind. Hans Moravec of Carnegie Mellon University is sure that after 2,050 most scientific discoveries will be made by artificial beings. Robots will develop their own cultures and goals. Their development and evolution will be much faster than biological development and evolution. Obsolete and unnecessary, humans could be condemned to extinction, but this perhaps will not be as terrible as it sounds(23): we could survive in "natural parks" or, even if our bodies are destroyed, computers could let our minds live in the form of conscious software. Moravec is not an isolated visionary. English philosopher Nick Bostrom also foretells the possibility of the supremacy of artificial intellects in this century. And Hugo de Garis, head of the Brain Builder Group at the Advanced Telecommunications Research Lab in Kyoto, Japan, is trying to make a hyper-intelligent electronic brain, applying Darwinian natural selection to electronic neural nets. He is sure, as are several other researchers, that before the end of this century there will be machines with much more intelligence than humans. He seems both scared and thrilled by this possibility: "My dream is to create 'artilects', or artificial intellects, with intellectual capacities many orders of magnitude above human levels."(24) He argues that the rise of artilects can perhaps represent the beginning of a huge planetary war between humans (those against versus those in favor of building them) or between humans and artilects themselves: "human beings will have the opportunity in the 21st century to build gods. […] The price we may have to pay may be a gigadeath war. […] That's what keeps me awake at night." He nonetheless thinks that we should continue in this direction: "I think humanity should build artilects-gods. To choose never to do such a thing would be a tragic mistake. […] I suspect that most of the advanced civilizations 'out there' […] have become artilects."

Weizenbaum, terrified by these visions, says the ideas of these radical supporters of post-human intelligences are "as dangerous as Hitler's *Mein Kampf*". Even without considering these science fiction scenarios, AI applications pose, no doubt, several crucial political and ethical questions about employment and worker rights, about social control permitted by technology, about democracy. On the other hand, some present applications of AI have already had positive impacts on society.

That's a funny conclusion: we don't know if "thinking" machines can really exist. We don't even know what thought, consciousness and intelligence mean. But while we debate the apple of discord – "Can machines think?" –, computers have bitten the apple of knowledge. Intelligent systems have spread all over the world. They are already radically modifying our lives and our vision of the world. The forbidden fruit of AI is, perhaps, already red and ripe. Like Sappho's apple.(25)

Notes and References

1. The most famous romance about the artificial being made by clay is Gustav Meyrinck's *Der Golem*, 1915 (an English version is *The Golem*, Dedalus; Riverside, 1995). Director Paul Wegener recorded a movie series about the monster, of which especially *Golem: How He Came Into the World* (USA, 1920, 1921) is famous. For the literature about robots, automata and artificial creatures, see also Y. Castelfranchi, O. Stock, *Macchine come noi. La scommessa dell'intelligenza artificiale*, Laterza, Roma-Bari, 2000-2002.
2. See, for example, L. Russo, *La rivoluzione dimenticata. Il pensiero scientifico greco e la scienza moderna*, Feltinelli, Milan, 1996.
3. Paracelsus was such a charismatic character that in English the adjective bombastic, used still today for "pompous", comes from his name.
4. Y. Castelfranchi, "Imaginando uma paleontologia da cultura científica", *Comciência*, n. 45, July 2003; http://www.comciencia.br/reportagens/cultura/cultura17.shtml.

5. Before Pascal's machine, it seems that Wilhelm Schickard built, in 1624, a "calculator clock" that maybe was able to perform the four operations. Unfortunately, it was destroyed in a fire.
6. A. M. Turing, "On computable numbers, with an application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society*, Ser. 2, Vol. 42, 1937; and Ser. 2, Vol. 43, 1937.
7. Shortly before Turing, and independently, Alonzo Church, in USA, also solved the problem, in a totally different way.
8. The history of computers is quite complex, and we don't enter here into details. See, for example, Castelfranchi & Stock (2000-2002), op. cit.
9. A. M. Turing, *Computing Machinery and Intelligence*, Mind, 1950, vol. 59, no. 236, pp. 433 – 460.
10. Of course, the definition of AI implies deep reflections about the definitions of what "intelligence", "consciousness" and "artificial" mean. We do not enter here into this debate. See, for example, L. Aiello, M. Mayer, *Invito all'Intelligenza Artificiale*, Franco Angeli, Milano.
11. C. Shannon, "Programming a Digital Computer for Playing Chess", *Philosophy Magazine*, n. 41, 1950.
12. See "Automated Intelligent Pilots for Combat Flight Simulation", in *AI Magazine*, spring 1999.
13. *Eliza* still lives in several places in the net. One place is: http://www-ai.ijs.si/eliza-cgi-bin/eliza_script.
14. M. Minsky, *The Society of Mind*, Simon and Schuster, New York, 1986.
15. B. Grosz et al., "Planning and Acting Together", in *AI Magazine*, winter 1999.
16. H. Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason*, Harper & Row, New York, 1972. See also H. Dreyfus, *What Computers "Still" Can't Do: A Critique of Artificial Reason*. Revised edition. MIT Press, Cambridge, 1992.
17. P. Benacerraf, "God, the Devil, and Godel", *Monist*, 51:9-32, 1967; J. Lucas, "Minds, machines and Godel", *Philosophy*, 36:112-127, 1961; R. Penrose, *The emperor's new mind: concerning computers, minds, and the laws of physics*, Oxford University Press, Oxford, 1989. J. Lucas, "Mind, machines and Godel: A retrospect", in P. Millican & A. Clark, eds, *Machines and Thought*, Oxford University Press, 1996
18. S. Feferman, "Penrose's Godelian argument", *Psyche* 2:21-32, 1996; D. McDermott, "Penrose is wrong", *Psyche* 2:66-82, 1996. D. McCullough, "Can humans escape Godel?", *Psyche* 2:57-65, 1996.

19. Gödel, it has to be said, had already understood perfectly well all implications of his theorems and even forecasted such discussions. In his final years, he wrote that either the human mind surpasses all possible machines, or "some problems in number theory exist that are undecidable for the human mind". See H. Wang, *From Mathematics to Philosophy*, London, 1997.
20. J. Searle, "Minds, Brains and Programs", in *The Behavioral and Brain Sciences*, 3, 1980
21. M. Boden, "Escaping from the Chinese Room", in *Computer Models of Mind*, Cambridge University Press, 1988. P.M. Churchland, & P.S. Churchland, "Could a machine think?", *Scientific American* 262(1):32-37, 1990. A very good on-line bibliography about AI, mind and consciousness is in http://jamaica.u.arizona.edu/~chalmers/online2.html.
22. The American Association for Artificial Intelligence provides a good on-line bibliography on AI and Ethics: http://www.aaai.org/AITopics/html/ethics.html.
23. C. Platt, "Superhumanism", *Wired*, 3 October 1995; H. Moravec, *Robot: Mere Machine to Transcendent Mind*, Oxford Univ. Press, 1999.
24. See, for example, R. Simpson, "The Brain Builder", *Wired*, 5 December 1997; or the on-line book by de Garis, *The Artilect War*, in http://www.cs.usu.edu/~degaris/artilectwar.html.
25. Acknowledgments: the English text was carefully revised and corrected by Prof. Robert Garner. I thank also, for useful discussions and contributions, Prof. Oliviero Stock, co-author with me of a book on the history of AI.