Artificial Intelligence («AI») is an interdisciplinary endeavour to make machines intelligent

Philippe-André Holzer

Specification

If we speak about machines and intelligence, we should be able to specify the meaning of these two terms.

There is a broad consensus in the scientific community that, among the machines we know, digital computers are the best candidates for obtaining intelligence; for this reason, practitioners of AI focus their efforts on making digital computers intelligent. We know how digital computers work. They are rule-driven manipulators of symbols, as specified by Alan Turing; they are realizations of Universal Turing Machines («UTMs»).[1] UTMs have characteristics that make them similar to intelligent human beings: they are good at mathematics; they are programmed with the help of (programming) languages and their in- and output is often language-like; furthermore, the distinction between hard- and software seems to parallel the distinction between body and mind in human beings. And there is sustained progress in our ability to build more reliable and faster computers and to program them in efficient ways.

To give a robust specification of intelligence or understanding is more difficult. Intelligence seems to be a characteristic we attribute without any substantial doubt to human beings, but which cannot be observed or measured in any direct way. What can be observed and measured is the actual behaviour of entities to which we ascribe intelligence. That is the reason why tests have been devised to measure intelligence. The idea is, then, to subject conveniently programmed computers to the tests by which we measure human intelligence; such tests may have to be enhanced to make sure they only measure intelligence and are not influenced by any other characteristics of the tested entities – but this is also a requirement when such tests are applied to human beings. If computers get high marks on such tests, we will ascribe intelligence to them. Alan Turing's proposal was to let a person have a teletyped dialogue with either another human being or a digital computer; when the person is no longer able to say whether the interlocutor is a computer or a human being, we should say that the computer actually thinking.[2] Such a way of dealing with the question of intelligence has the advantage that we do not have to wait for the solution of the many scientific and philosophical questions involved in the determination of what intelligence is. But this advantage comes at a price. It is a shift from intelligence to the test of intelligence; we assume that human beings who are successful in the test have intelligence, but we do no really know the nature of the characteristic we ascribe to the computer.

The result of our preliminary specification is to say that AI is the effort to make digital computers do what requires intelligence when human beings do it.

Some history

The efforts in AI began with the development of digital computers. The term was first used in the 1950s. Expectations were high, but soon problems began to show up. The development of viable AI applications on digital computers is much less easy than expected. That is why in the history of AI, times of great enthusiasm and promise are followed by periods in which the problems seem to be overwhelming (the so- called «AI winters»).[3] As it became clear that the goal of implementing intelligence in a computer would not be reached soon by one single project, the field of AI has di-

vided into a bundle of projects and strategies. These have to do with sense perception (artificial vision, pattern recognition, interpretation of sensor data), natural language (speech recognition, language understanding and generation), reasoning (proof of mathematical theorems, problem solving, everyday situations), learning, planning, creativity (computers as contemporary artists), decision taking and, finally, with general intelligence. All these activities turned out to be much more difficult to realize on computers than anticipated. Different research and programming strategies have been proposed. While in some projects, contents are represented in a numeric or linguistic (that is: symbolic) way, in others, the goal is rather to implement the processes that make human symbolic representation possible, namely feedback processes and the activities going on in our nervous system, ranging from simple neurons to the integrated structure of a whole human brain. At some point, these different methodologies will have to be convincingly integrated.

Philosophical questions

The successes of AI sometimes cause delusions: if a given AI problem has been solved, it often turns out that the solution is not any longer considered to be part of AI. Take the task of beating the best human chess players at chess. This task has been successfully solved and, at first sight, this seems to be impressive. After all, successful chess playing may be seen as an obvious indicator of intelligence and the ability to think strategically. If, on closer inspection, however, it turns out that the computer has reached its goal by making a very efficient and fast search through a database of chess moves, we are less impressed by it – this type of activity seems to be far less intelligent than the strategic thought of a human player. It remains true, however, that the computer somehow outdid the competitors, because no human being would be able to search through a database as efficiently and as fast as a digital computer.

As long as you only consider the result (in our example: the chess games won), you are impressed with the output of the computer and might have the tendency to ascribe intelligence to it. As soon as you come to know the ways computers are programmed and reach their output, however, it is much more likely that you will think that the computer is only formally manipulating symbols; then, you might have the tendency to think that the computer only simulates intelligence instead of literally possessing that characteristic. Whoever is convinced that we should ascribe genuine intelligence to computers will tend to tell us that we should consider human beings and computers as black boxes and judge them only on their respective in- and outputs. Whoever is critical of any such easy attribution to the computer of thought will point to the many ways in which human thinking differs from what goes on in digital computers.

Do computers actually think or do they just simulate thinking? The former position is known as «strong AI», the latter as «weak AI». Joseph Weizenbaum has presented an argument for the weak version by showing how easy it is to program a computer in such a way that in a teletyped dialogue it acts like a psychotherapist following the approach of Carl Rogers. If you judge according to the behaviour of the computer, you are impressed because it acts like an understanding and rather sympathetic person (and several persons actually took it as such). If you know how the computer is programmed, however, you see that it follows certain dialogue strategies without having any understanding of what is said (better: typed) during the conversation.[4]

The most influential philosophical argument against strong AI is the thought experiment of the «Chinese Room» by John Searle.[5] According to him, intelligence and thinking are characteristics and activities of living human beings; as far as

we know, they are only possible through the causal powers of the human brain. We have no reason to believe that they could be embodied in non-biological entities, for example in digital computers. Such machines simply lack the causal powers in question.

Searle highlights the fact that human beings are able to calculate, but also to perform and to have many other psychological acts and states, such as natural language understanding. We know what it means for us to calculate: it means to perform manipulations of symbols according to formally-defined rules. Let us think of a human person in a room, a native English speaker who does not know Chinese, at a desk with rulebooks and paper (for writing down intermediary results). This room is connected to the outside world by a hole, through which symbol tokens can be handed from the outside into the room (the input) and through which the person in the room can give back similar tokens (the output). If tokens are given as input, the person in the room looks up in a book of rules (written entirely in one language, let us say in English) which other tokens should be handed out as output. These rules make only reference to purely formal characteristics of the symbol tokens. This arrangement would count as a realization of a Universal Turing Machine.

Let us assume, further, that the rules are written in such a way that if Chinese characters are given as input, the person in the room will give other Chinese characters as output, so that a sustained dialogue in Chinese is possible. Let us finally assume that the person in the room is well-trained and fast, so that the time between input and output is reasonably short. Then, for a Chinese person outside the room, it would seem that the room is fluent in Chinese, as the answers make perfect sense in Chinese. In this way, «the Chinese room» would pass the Turing Test in Chinese.

But the person in the room does not know Chinese. The rules are written in English, and the symbol tokens bearing the Chinese characters are identified exclusively by formal descriptions in English. The person in the room consciously performs the formal symbol manipulations (and therefore calculates according to the requirements for UTMs) by which fluency in Chinese is simulated. But this person – on the grounds of these consciously performed activities – has no idea what is done with the symbol manipulations performed.

Searle's article has been widely discussed, and many answers mirror the different approaches within AI. Some say that the knowledge of Chinese has to be ascribed not to the person in the room, but to the system comprising the person and the room; others think true AI requires a causal connection between computer and environment through sensors and actors; others argue that true AI will come into being through the digital replication of the signals present in our nervous system or in the whole brain. But these responses do not give an answer to the fundamental question: are the computational activities present in digital computers sufficient for the existence of genuinely intelligent, intentional, mental acts or states, like the ones we find in human beings? A positive answer is often presupposed and not argued for.

Traditional philosophical reflection on the acts of human beings provides us with a distinction that still seems to be useful in dealing with AI: the distinction between transitive and immanent action, between making (something happen) and doing.[6] A transitive action changes the world in a very precise way: if you make something happen (you plough a field, for example), that something may be described, measured, depicted, even filmed – it can be clearly decided what you did to make it happen. But in an immanent action, what is changed is not the world, but the human being who does it: this happens when you sense, see, hear, understand, think, learn, answer and remember something. The outcome of such an action is not measurable and cannot be filmed. Nevertheless, it is real and characterizes you as a human being with knowledge, experience and perception. In contemporary philosophy, such immanent acts and states are characterized as intentional, which implies a certain way of being conscious.

This distinction between transitive and immanent acts makes it understandable that the actions of *machines* or *computers* can be precisely specified, whereas this is not the case with *intelligence* or *thinking*. The latter two are characterized by intentionality and consciousness; to try to reduce them to some non-intentional and non-conscious counterparts runs against the spirit of science.[7]

Conclusion

Immanent actions and states of human beings are characterized by intentionality and consciousness. Genuine intelligence should not be attributed

to entities without intentionality and consciousness. This does not exclude a certain extension of intentionality outside human beings; with tools, machines and other artefacts, human beings have been very successful at «exporting» intentionality into the environment, such as into books and digital computers. If used in a proper way, tools and books «make perfect sense» and are often said to be «smart» – but their sense and smartness is more on the side of those who use them properly than on their own side. Outside this proper use, they do not seem to be smart or intelligent on their own.

Properly used, they outperform human beings in specific transitive tasks[8] – if this were not the case, it would be pointless to develop and use them. This gives special power and responsibility to those who develop and use them.

We should keep in mind that the promise of an intelligent machine is especially appealing to the military: systems able to interpret pictures, to understand and translate natural language, to ponder different aspects of a problem and to take decisions could be used for espionage, surveillance, threat assessment and automated response in a fraction of the time human beings would need. Progress in AI is deeply linked with military applications thereof.

NOTES

[1] Cf. Alan Matison Turing's abstract specification of a universal digital computer in «On computable numbers, with an application to the Entscheidungsproblem», in: *Proceedings of the London Mathematical Society,* Second Series 42 (1937), 230–265.

[2] This is the so-called «Turing Test» as put forward in «Computing Machinery and Intelligence», in: *Mind* 59 (1950), 433–460.

[3] For the history of AI and the main projects and approaches see Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements,* New York: Cambridge University Press, 2010.

[4] See Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation,* San Francisco: Freeman, 1976.

[5] «Minds, brains, and programs», in: *Behavioral and Brain Sciences* 3 (1980), 417–457.

[6] For Evandro Agazzi, this distinction is very important for a critical reflection on the problems of AI; see, for example, «Operazionalità e intenzionalità: L'anello mancante dell'intelligenza artificiale», in: *Intelligenza naturale e intelligenza artificiale: Contributi al XLIII Convegno del Centro di Studi Filosofici di Gallarate (Aprile 1988),* edited by Salvino Biolo, Genova: Marietti, 1991, 1–13.

[7] Evandro Agazzi, «Reductionism as Negation of the Scientific Spirit», in: *The Problem of Reductionism in Science,* Dordrecht: Kluwer, 1991, 1–29.

[8] According to Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies,* Oxford: Oxford University Press, 2014, AI could radically outperform human beings in the future and take over a considerable part of the universe.