MONOGRAPH
MÈTODE SCIENCE STUDIES JOURNAL (2018). University of Valencia.
DOI: 10.7203/metode.9.11145
ISSN: 2174-3487 / eISSN: 2174-9221.
Article received: 12/12/2017, accepted: 23/07/2018.

TOWARDS ARTIFICIAL INTELLIGENCE

ADVANCES, CHALLENGES, AND RISKS

RAMON LÓPEZ DE MÁNTARAS

This text contains some reflections on artificial intelligence (AI). First, we distinguish between strong and weak AI, as well as the concepts related to general and specific AI. Following this, we briefly describe the main current AI models and discuss the need to provide common-sense knowledge to machines in order to advance towards the goal of a general AI. Next, we talk about the current trends in AI based on the analysis of large amounts of data, which has recently allowed experts to make spectacular progress. Finally, we discuss other topics which, now and in the future, will continue to be key in AI, before closing with a brief reflection on the risks of AI.

Keywords: strong artificial intelligence, weak artificial intelligence, common-sense knowledge, deep learning.

INTRODUCTION

The ultimate objective of the field of artificial intelligence (AI), the creation of a machine with a general intelligence like that of humans, is one of the most ambitious scientific goals ever set. The difficulty of this is comparable to other great scientific objectives such as explaining the origin of life or of the universe or understanding the structure of matter. Over the last few centuries, this desire to build intelligent machines led us to invent models or metaphors for the human brain. For instance, in the seventeenth century, Descartes wondered if a complex mechanical system consisting of gears, pulleys, and tubes could, in principle, emulate thinking. Two centuries later, this metaphor was applied to telephone systems because their connections seemed to be like a neural network. Currently, the dominant AI model is based on digital computers and computation, as we discuss in this article.

«AI IS THE SCIENTIFIC FIELD DEVOTED TO TRYING TO VERIFY WHETHER A COMPUTER CAN BEHAVE WITH GENERAL INTELLIGENCE OR NOT»

WEAK VERSUS STRONG ARTIFICIAL INTELLIGENCE

Allen Newell and Herbert Simon formulated the hypothesis according to which every physical symbol system has the necessary and sufficient means for intelligent action (Newell & Simon, 1976). According to this hypothesis, while human beings show intelligent behaviour, we are also physical symbol systems. We should be clear on what Newell and Simon referred to. A physical symbol system consists of a set of entities known as symbols which, via their relationship to each other, can be combined to form larger structures – like atoms combining to form molecules – and can be transformed using several procedures. These procedures can create new symbols, create and modify the relationships between them, store them, compare two of them to see if they are the same or different, etc. These symbols are physical because they have a physical-electronic substrate (in the case of computers) or a physical-biological substrate (in the case of humans). Indeed, symbols in computers are created using digital electronic circuits; in humans, they use neural networks. In short, according to the hypothesis of the physical symbol system, the nature of the substrate (electronic circuits or neural networks) is not important if it allows the system to process symbols. Let us not forget that this is a hypothesis and thus, its verification or refutation must use the scientific method. Artificial intelligence is precisely the scientific discipline devoted to trying to verify this hypothesis in the context of computers; that is, to verify whether a properly programmed computer can behave with general intelligence or not.

It is important for such an intelligence to be general, instead of specific, because that is the type of intelligence humans have. Displaying specific intelligence is very different. For example, master-level chess programmes cannot play checkers. The same computer requires a different programme to play checkers; it cannot use the fact that it can play chess to adapt so that it can also play checkers, while human chess players can take advantage of their knowledge of chess to play checkers. Artificial intelligence that shows only very specific intelligence is related to what we know as «weak AI», in contrast with «strong AI», which, incidentally, is the one Newell, Simon, and other forefathers of AI referred to.

«ALL THE ADVANCES MADE SO FAR ARE MANIFESTATIONS OF SPECIFIC AND WEAK AI»

The philosopher John Searle was the first to introduce the distinction between weak and strong AI in a paper published in 1980 that criticised artificial intelligence (Searle, 1980) that raised, and still raises, many doubts. Strong AI would imply a properly programmed computer that does not emulate a mind, but rather «is a mind», so it should be able to think as a human does. In his paper, Searle tries to prove that strong AI is impossible.

At this point, we should be clear that general AI and strong AI are not the same thing. There is a connection, of course, but only in one direction: that is, any strong AI must necessarily be general, but general AIs that are not strong can exist, meaning that they simulate the ability to show general intelligence without being real minds.

On the other hand, and according to Searle, weak AI would consist of programmes which perform specific tasks. In certain fields it has been amply demonstrated that the ability of computers to perform specific tasks like searching for the solutions to logical formulas with many variables and other decision-making tasks is superior to that of humans. Weak AI is also connected with the formulation and proof of hypotheses on aspects related to the mind (for example, the ability to reason deductively, to learn inductively, etc.) via the construction of programmes that carry out these functions, even if the processes they use are different from those of our brains. It is clear that all the advances made so far in the field of AI are manifestations of specific and weak AI.

THE MAIN MODELS IN ARTIFICIAL INTELLIGENCE

Until very recently, the leading AI model was the symbolic one. This model is rooted in the physical symbol system hypothesis. It is still very important and is currently considered the «classic» AI model. It is a top-down model based on logical reasoning and heuristics as pillars for problem solving, without the need for the intelligent system to be part of a body or be in a real environment. That is to say, the symbolic AI operates with abstract representations of the real world, modelled using representation languages based mainly in mathematical logic and its extensions. For this reason, the first intelligent systems mainly solved problems that did not require direct interaction with the environment, like proving simple mathematical theorems or playing chess. This does not

Carnegie Mellon University

In the 1970s, Allen Newell and Herbert Simon suggested that any physical symbol system – whether they were physical-electronic in the case of computers or physical-biological in the case of humans – has the necessary means to carry out intelligent actions. The picture shows professors Simon (left) and Newell (right) working on chess software at the end of the 1950s, at Carnegie Mellon University in Pittsburgh (USA).

IBM

Human intelligence is of the general type, while the intelligence of grandmaster-level chess software, like the Deep Blue computer that won against Kasparov in 1997, is of the specific type. This means that chess-playing computers are unable to use their knowledge to play checkers, for example. The picture shows the IBM team that developed Deep Blue in a photograph from 1996.

mean that symbolic AI cannot be used to programme the reasoning module of a physical robot situated in a real environment, but in the first years of AI there was no knowledge representation or programming language that could do it efficiently. Currently, symbolic AI is still used to prove theorems or play chess but it is also now used for applications that require environmental awareness and action, such as autonomous robots that learn and make decisions.

«IT HAS BEEN AMPLY DEMONSTRATED THAT THE ABILITY OF COMPUTERS TO PERFORM SPECIFIC TASKS IS EVEN BETTER THAN THAT OF HUMANS»

At the same time symbolic AI was being created, researchers also started to develop a bioinspired AI called connectionist AI. Contrary to symbolic AI, this is a bottom-up model based on the hypothesis that intelligence emerges from the distributed activity of many interconnected units processing information at the same time. In connectionist AI, these units are very approximate models of the electric activity of biological neurons. McCulloch and Pitts (1943) proposed a simplified neuron model in accordance with the idea that a neuron is essentially a logical unit. This model is a mathematical abstraction with inputs and outputs corresponding respectively to dendrites and axons. The value of the output is calculated depending on the result of the weighted sum of the inputs, so that if the sum exceeds a pre-established threshold, then the output is 1; otherwise, the output is 0. Connecting the output of each neuron with the input of others forms an artificial neural network. Based on what was already known about reinforcing the synapses between biological neurons, it was discovered that these artificial neural networks could be trained with functions that connected inputs and outputs by adjusting the weights of the connections between neurons. Therefore, experts thought that the connectionist approach would be more appropriate than the symbolic one to model learning, cognition, and memory. However, intelligent systems based on connectionism do not need to be part of a body or be situated in a real environment either; from this perspective, they have the same limitations as symbolic systems.

On the other hand, 90% of brain cells are not neurons but rather, are glial cells which do not only regulate the functioning of neurons but also have an electric potential. They generate calcium waves which allow intercommunication between each other, which indicates that they also play a very important role in cognitive processes. Nevertheless, no connectionist model includes these cells; therefore, in the best case, they are incomplete. This suggests that the prediction that the so-called singularity, the moment when artificial superintelligences replicate a brain which far surpasses human intelligence, will be reached within twenty years' time has little foundation.

Another bioinspired model, also unembodied and compatible with the physical symbol system hypothesis, is evolutionary computation. The biological success of having evolved complex organisms led some researchers in the early sixties to consider the possibility of imitating evolution so that, using an evolutionary process, computer software would automatically improve the solutions to the problems for which it had been created. The idea is that these programmes produce, thanks to the mutation and crossbreeding operators of the «chromosomes» modelling them, new generations of modified programmes that offer better solutions than the ones from previous generations. Given that the goal of AI consists in the development of programmes that can produce intelligent behaviour, experts supposed that evolutionary programming could be used to find such programs from among the spectrum of possible programmes. However, the reality is much more complex, and this approach has many limitations, although it did produce excellent results, particularly in the resolution of optimisation problems.

One of the strongest criticisms of these unembodied models is that an intelligent agent would need a body to directly experience its environment instead of being provided with abstract descriptions of that environment codified in a knowledge-representation language. Without a body, these abstract representations have no semantic content. However, through direct interaction with the environment, an embodied agent can connect the signals received through its sensors with symbolic representations generated from what it has sensed.

«ONE OF THE STRONGEST CRITICISMS OF UNEMBODIED MODELS IS THAT AN INTELLIGENT AGENT WOULD NEED A BODY SO THAT IT COULD DIRECTLY EXPERIENCE ITS ENVIRONMENT»

In 1965, the philosopher Hubert Dreyfus published a paper titled «Alchemy and artificial intelligence» (Dreyfus, 1965), in which he claimed that the ultimate goal of AI – meaning strong, general AI – was as unattainable as the goal of the seventeenth-century alchemists who wanted to transform lead into gold. Dreyfus argued that our brain processes information globally and continuously, while a computer uses a finite and discrete set of deterministic operations; that is, it applies rules to a finite set of data. In some sense, this argument is like Searle's, but in subsequent papers and books, Dreyfus used a different argument based on the essential role that the body plays in intelligence (Dreyfus, 1992). Therefore, he was one of the first to advocate the need for intelligence to be implemented as part of a body that could interact with the world. The main idea is that the intelligence of human beings is derived from the fact that they are situated in an environment with which they can interact. In fact, this need for embodiment is based

Softbank Robotics

One of the criticisms of non-corporeal artificial intelligences is that an intelligent agent needs a body to be able to experience the world directly. The picture shows the Romeo humanoid robot developed by SoftBank Robotics.

«NO MATTER HOW SOPHISTICATED THEY BECOME, MACHINES' ARTIFICIAL INTELLIGENCES WILL ALWAYS BE DIFFERENT FROM OURS»

IBM

The availability of enormous amounts of data and access to high-performance computing to analyse them has enabled the development of new artificial intelligence systems such as Watson, which can answer natural language questions. According to its creator, IBM, Watson can learn from every experience.

on Heidegger's phenomenology, which emphasises the importance of the body in his needs, desires, pleasures, sorrows, the way the body moves, acts, etc. According to Dreyfus, AI would have to model all these aspects to achieve the ultimate objective of creating a strong AI. In other words, Dreyfus does not completely deny the possibility of strong AI, but he claims that it is not possible to achieve by means of classical unembodied AI methods.

«COMMON SENSE IS THE FUNDAMENTAL REQUIREMENT FOR AN AI TO BE LIKE HUMAN INTELLIGENCE IN TERMS OF GENERALITY AND DEPTH»

DO THE ADVANCES IN SPECIFIC ARTIFICIAL INTELLIGENCE BRING US CLOSER TO GENERAL ARTIFICIAL INTELLIGENCE?

Virtually all AI efforts have focused on building specialised artificial intelligences, and the successes of the last sixty years, and particularly the last decade, are very impressive, mainly thanks to the combination of two elements: the availability of enormous amounts of data and access to high-performance computing to analyse them. Indeed, the success of systems such as AlphaGo (Silver et al., 2016), Watson (Ferrucci, Levas, Bagchi, Gondek, & Mueller, 2013) and advances in autonomous vehicles have been possible thanks to this ability to analyse large amounts of data. However, we have not progressed towards achieving a general AI. In fact, possibly the most important lesson we have learned over the course of sixty years of AI is that what seemed hardest in the past (diagnosing diseases or playing chess at the highest level) has turned out to be feasible and what seemed easiest (understanding the deep meaning of language or interpreting a visual scene) remains out of reach.

We must look for the explanation of this apparent contradiction in the difficulty of providing machines with common-sense knowledge. Common sense is the fundamental requirement for an AI to be like human intelligence in terms of generality and depth. Common-sense knowledge is the fruit of experiences obtained through interactions with our environment. Without this knowledge, it is impossible to deeply understand language or profoundly interpret the perceptions of a visual system, among other limitations. The most complicated skills to achieve are those that require interacting with unrestricted and not previously prepared environments. Designing systems with these capabilities requires integrating developments from many subfields of AI. In particular, we need knowledge-representation languages to codify information about, among others, many different types of objects, situations, actions, and the properties and interconnections between them.

We also need new algorithms that use these representations to respond robustly and efficiently to questions on virtually any topic. Finally, because these systems will need to know a virtually unlimited number of things, they must be able to learn new knowledge continuously throughout their existence. In short, designing systems that integrate perception, representation, reasoning, action, and learning is essential. Only by combining these elements within integrated cognitive systems we will be able to start building general AI.

THE RECENT-PAST AND SHORT-TERM FUTURE OF ARTIFICIAL INTELLIGENCE

Among future activities, I believe that the most important research topics will continue to depend on massive data-driven AI, that is, they will take advantage of large amounts of data and process them with increasingly-powerful hardware in order to discover how they relate to each other, to detect patterns and learn using statistical approaches such as deep-learning systems (Bengio, 2009). However, in the future, systems based on analysing enormous amounts of data will have to include modules to explain how the results and conclusions they propose were reached, because any intelligent system must necessarily be able to explain itself. The explicability will allow to understand how the system works and evaluate its reliability. On the other hand, explicability is also needed to be able to correct potential programming errors and detect whether the training data was biased.

We also need to know if the results are correct for the right reasons or just because of coincidences in the training data. Therefore, one of the most important research topics in the field of deep learning is the design of interpretable approaches to these complex learning systems. A possible approach would consist of not only training the deep-learning system but also using the same data set to train another system, emulating the deep-learning one, using a simple and transparent representation.

Another current research topic is the verification and validation of the software that implements learning algorithms. This is especially important in high-risk applications such as the autopilot in autonomous vehicles. In these cases, we need a methodology to verify and validate high levels of precision for these machine learning systems. An idea that is currently being explored is called adversarial learning, which requires training a second AI system to attempt to «break» the learning algorithm by attempting to find its weak points. For instance, in the case of visual recognition, the adversarial system generates images that try to fool the learning system into making the wrong decision.

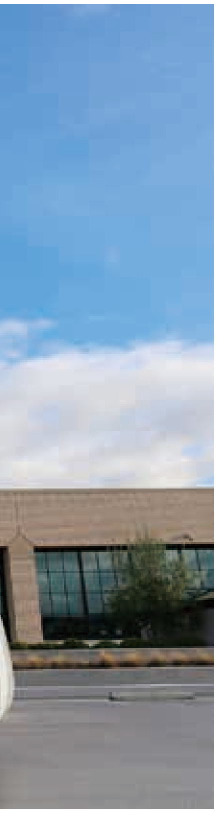
Waymo

The development of technologies related to artificial intelligence means that we must analyse their potential risks. For instance, in the case of autonomous vehicle piloting, we need a methodology for verifying and validating high levels of precision in their machine-learning systems.

«EVEN IF IT WERE POSSIBLE TO DEVELOP COMPLETELY RELIABLE SOFTWARE, PROGRAMMERS MUST BE AWARE OF ETHICAL ISSUES INVOLVED WHEN DESIGNING IT»

■ OTHER KEY TOPICS IN ARTIFICIAL INTELLIGENCE

Other areas of AI that will continue to be the focus of extensive research efforts are multiagent systems, action planning, experience-based reasoning, artificial vision, human machine multimodal communication, humanoid robotics, social robotics, and the new developmental robotics trends, which may be crucial to solve the problem of allowing machines to acquire common sense. We will also see significant advances thanks to biomimetic approaches which reproduce animal behaviour in machines. Some biologists are interested in creating an artificial brain that is as complex as possible because they think it is a good way to understand the organ better, while engineers, on the other hand, look for biological insights to create more effective designs.



Other important areas for AI, and especially for robotics, are materials science and nanotechnology; for example, for the development of artificial muscles, artificial cartilages, and sensory systems such as artificial skins.

Regarding the applications, some of the most important ones are still those related to the Internet, video games, and autonomous robots (especially autonomous vehicles, social robots, planet exploration robots, etc.). Environmental and energy-saving applications will also be important, as well as economy and sociology applications.

Finally, the art applications of AI will significantly change the nature of the creative process. Computers are no longer just tools to help creators, they are starting to become creative agents. This has given rise to a new and very promising area in artificial intelligence called computational creativity, which has produced very interesting results (Colton, López de Mántaras, & Stock, 2009; Colton et al., 2015; López de Mántaras, 2016) in music and plastic and narrative arts, among other creative activities.

■ THE RISKS OF ARTIFICIAL INTELLIGENCE, A FINAL REFLECTION

No matter how intelligent future artificial intelligences will be, particularly general AIs, they will never be like human intelligence, because, as I have argued, the mental development required by any complex intelligence depends on the interactions with its environment, which, in turn, depend on the body, especially its perceptual and motor systems. The fact that the socialisation and education of machines, if any, will be different from ours further emphasises that, no matter how sophisticated they become, their intelligence will always be different from ours. The fact that these intelligences are alien to human intelligence and, therefore, alien to human values and needs, should make us reflect on the possible ethical limitations to the development of artificial intelligence. In particular, I agree with Weizenbaum (1976) when he says that no machine should make completely autonomous decisions or give advice that requires, among other things, the wisdom, which is the product of human experience and human values.

AI is based in complex programming; therefore, it will necessarily contain errors. But even if it were possible to develop completely reliable software, programmers must be aware of ethical issues involved when designing it. These ethical aspects lead many AI experts to discuss the need to regulate their development. But, aside from such regulation, we must educate citizens on risks related to intelligent technologies, providing them with the skills required to control these technologies, rather than be controlled by them. This education process must start at school and continue in higher education. We especially need science and engineering students to receive ethics training so they can better comprehend the social implications of the technologies that they will very probably develop. Only if we invest in education will we achieve a society that can take advantage of intelligent technologies while minimising its risks. ⊕

REFERENCES

Bengio., Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127. doi: 10.1561/2200000006

Colton, S., Halskov, J., Ventura, D., Gouldstone, I., Cook, M., & Pérez-Ferrer, B. (2015). The Painting Fool sees! New projects with the automated painter. In *International Conference on Computational Creativity (ICCC 2015)* (pp. 189–196). Utah, UT: Brighma Young University.

Colton, S., López de Mántaras, R., & Stock, O. (2009). Computational creativity: Coming of age. *AI Magazine, 30*(3), 11–14. doi: 10.1609/aimag.v30i3.2257

Dreyfus, H. L. (1965). *Alchemy and artificial intelligence*. Santa Monica, CA: RAND Corporation.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: MIT Press.

Ferrucci, D. A., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. T. (2013). Watson: Beyond Jeopardy! *Artificial Intelligence*, *199*, 93–105. doi: 10.1016/j.artint.2012.06.009

López de Mántaras, R. (2016). Artificial intelligence and the arts: Toward computational creativity. In *The next step: Exponential life* (pp. 100–125). Madrid: BBVA.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133. doi: 10.1007/BF02478259

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. doi: 10.1145/360018.360022

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. doi: 10.1017/S0140525X00005756

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van den Driessche, ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. doi: 10.1038/nature16961

Weizenbaum, J. (1976). *Computer power and human reasoning: From judgment to calculation*. San Francisco, CA: W. H. Freeman and Co.

Ramon López de Mántaras. Research Professor and Director of the Artificial Intelligence Research Institute at the Spanish National Research Council (IIIA-CSIC, Bellaterra, Spain). He holds a PhD in Physics from the Paul Sabatier University in Toulouse, a Master's Degree in Computer Science from the University of California, Berkeley, and a PhD in Computer Science from the Polytechnic University of Catalonia. He is also a numerary member of the Institute of Catalan Studies. He currently researches reasoning by analogy, machine learning techniques for humanoid robots, and artificial intelligence applied to music, and has published around 300 scientific papers in these fields. In 2017, he published the popular science book *Inteligencia artificial* within the «Qué sabemos de» collection (Los Libros de la Catarata). ✉ mantaras@iiia.csic.es