Briefing
January 2018

Understanding artificial intelligence

SUMMARY

Artificial intelligence (AI) systems already permeate daily life: they drive cars, decide on mortgage applications, translate texts, recognise faces on social networks, identify spam emails, create artworks, play games, and intervene in conflict zones. The AI revolution that began in the 2000s emerged from the combination of machine learning techniques and 'big data'. The algorithms behind these systems work by identifying statistical correlation in the data they analyse, enabling them to perform tasks for which intelligence is required if a human were to perform them.

Nevertheless, data-driven AI can only perform one task at a time, and cannot transfer its knowledge. 'Strong AI', able to display human-like intelligence and common sense, and which might be able to set its own goals, is not yet within reach. Despite the fears portrayed in film and TV entertainment, the idea of a 'superintelligence' able to self-improve and dominate humans remains an esoteric possibility, as development of strong AI systems is not predicted for a few decades or more, if indeed development ever reaches this stage.

Nevertheless, the development of data-driven AI systems implies adaptation of legal frameworks on the collection, use and storage of data, due to privacy and other issues. Bias in data supplied to AI systems can also reproduce or amplify bias in the decisions they make. However, the key issue remains the level of autonomy given to AI systems to make decisions that could be life-changing, keeping in mind that they only provide recommendations, that they do not understand the tasks they perform, and that there is no way to know how they reach their conclusions. AI systems are expected to impact society, especially the job market, and could increase inequalities.

To counter the abuse of probabilistic prediction and the risks to privacy, in April 2016 the European Parliament and the Council of the EU adopted the General Data Protection Regulation. The European Parliament also requested an update of the Union legal framework on robotics and AI in February 2017.

In this briefing:

EPRS | European Parliamentary Research Service
Author: Vincent Reillon
Members' Research Service

The dawn of artificial intelligence

In 1936, English mathematician Alan Turing proposed the concept of the 'Turing machine', a model of computation that triggered the development of informatics and computers. In 1950, Turing published a paper on 'Computing machinery and intelligence' that is often referred to as the origin of modern artificial intelligence – that is to say the capacity for a machine to display human-like capacities such as reasoning, learning, planning and creativity.[1]

Symbolic artificial intelligence

Artificial intelligence began developing in the 1960s around the idea that it should be possible to deconstruct intelligent human behaviours as a succession of logical rules, transcribed in algorithms, which machines could follow to display intelligent behaviour. The information given to the machine should then be transformed into symbols (graphs, logic formulas) that the computer could manipulate using a set of rules. This was a top-down approach, deconstructing a given behaviour into a set of smaller problems and a knowledge-based approach requiring a symbolic description of the world.

This approach led to the design of expert systems that reproduce the cognitive steps of a human expert to solve problems. An expert system is composed of a knowledge base that represents facts about the real world, and an inference engine that applies a set of logical rules to deduce new knowledge from the knowledge base. Expert systems are able to provide support in a large array of tasks, from controlling system behaviour to providing diagnosis to supporting decision-making.

The failure of symbolic reasoning

Translating knowledge into symbols and defining all the rules a machine would need to perform an assignment quickly became an overwhelming task. The programmers needed to consider all the possibilities that the machine might encounter so that everything it needed to operate was fully described. In addition, this approach required considerable computing power and the necessary hardware was lacking at the time. Moreover, research in neurosciences began to show that the human brain worked in a different way.

These issues led to episodes during which the initial enthusiasm for artificial intelligence waned and the associated funding for the field was cut. These periods, at the beginning of the 1970s and the end of the 1980s, are referred to as 'AI winters'. Despite the use of expert systems, symbolic AI appeared a dead end for the evolution of the field.

In 1997, Deep Blue's chess game victory against Garry Kasparov sparked a new interest in AI. Nevertheless, Deep Blue was based on an algorithm analysing millions of possibilities per second and selecting the most promising move, a 'brute force' technique supported by extensive computing power that merely appeared to display intelligence. A new approach was needed in order for AI to develop further.

Data-driven artificial intelligence

At the beginning of the 2000s, a new wave of AI would emerge from the combination of two elements: algorithms that allow machines to learn and the large quantity of data produced by the development of the digital world.

Machine learning techniques

Learning is one of the key features of human intelligence. In the field of artificial intelligence, learning is understood as the capacity to utilise experience to improve a behaviour. Neuroscience has shown that human cognitive capacities are based on the activation of complex networks of neurons in the brain. These neural networks are able to store information and knowledge and consequently provide learning abilities.

Inspired by this process, programmers created artificial neural networks (ANN). In an ANN, a large number of units (artificial neurons) are connected with each other to create a complex network of interactions with different layers. When given an input signal, the network produces an output signal that results from the interactions in the network. The key aspect of ANN is that the program is able to modify the interactions in the network until the given input produces the expected output, providing the network with the ability to be trained and to learn.

The ANN's knowledge is stored in the network itself, in a similar way to current understanding regarding the functioning of the human brain. Multiplying the layers of ANN and coupling different machine learning techniques led to what are known as 'deep learning' techniques – widely used for different tasks. Because of the large amount of data involved in the processes of machine learning, this new generation of artificial intelligence is referred to as data-driven AI.

In order to train these coupled networks, a large quantity of data is required. If the objective is to teach a machine to recognise pictures with cats, the machine is fed with thousands of pictures, including pictures with cats. Any time a picture is presented as an input to the network, it will provide an output: 'cat' or 'not cat'. If the output is correct, the network will strengthen its internal interactions. If the output is incorrect, it will modify its interactions to take this information into account. After analysing hundreds of thousands of pictures, the ANN interactions is set – in this case, the program will be able to recognise pictures with cats. It is then able to provide the expected output when presented with new pictures.

Machine learning can be supervised, or not. In the case of supervised learning, the machine is trained to realise a specific task, such as to recognise cats in pictures. This requires labelling a large quantity of data, thousands of pictures, containing cats or not following this example. It also implies that the machine is controlled to see if it provides the right answer for each picture it analyses during the training. Supervised training is used for tasks requiring the classification of information. In unsupervised training, the program is not given any task and data remain unlabelled. The program is free to find its own correlations in the data. Learning from the data, the machine will create clusters in the data provided, and provide association rules that connect different variables in the data. This is used for example to define clusters of customers.

Key aspects of machine learning techniques

The learning capacity of machines is based on the ability of the algorithm to find statistical correlations in the data it analyses, that is to say interdependence of variables in the data. In the example of the cat pictures, the program decomposes the data (i.e. each picture) in a series of variables. The training of the neural networks will progressively select the variables whose values correlates better with the pictures being one of a cat.

The program therefore has no idea what a cat is. The symbolic AI approach would have consisted of explaining to a program what a cat is, so that it could recognise an image of a cat. This would have entailed programming the machine to understand what a leg is, what a tail is, what fur is, and that a cat had four legs, a tail and could exhibit different fur patterns, before the program could recognise an image of a cat. Symbolic AI would therefore render this task both complicated and heavy in terms of the computing power and time required. Machine learning techniques bypass these barriers by using machines' ability to analyse huge quantities of data to find statistical correlations (a task at which the human mind does not excel). The advantage of data-driven AI is that machines can perform tasks that would be impossible or too complicated to program.

However, a fundamental aspect of machine learning techniques is that there is no way to know how the machine reaches its conclusion, how it makes its decision on a given task. In the example of images of cats, it is not possible to know which variables the program uses in the description of a picture to decide whether or not the image contains a cat. In symbolic AI the concept of 'explainability', the ability to explain how a system reached its conclusion, was central. In data-driven AI 'explainability' no longer matters. Only the result counts: what a machine can do, not how it does it. This also raises the question of determining whether data-driven AI systems really do what is expected of them. In the same way neurologists explore the way the brain works, researchers are currently analysing how AI systems proceed to reach their decisions.

Applications of data-driven artificial intelligence

The machine learning approach has led to a wide range of successes in various fields:

- Automatic translation, provided for example by Google Translate, DeepL or SYSTRAN;
- Speech recognition and interpretation, such as the example between English and Chinese demonstrated in November 2012 by Rick Rashid of Microsoft;
- Face recognition systems used in criminal investigations or to unlock a smartphone;
- Machines playing games: since Deep Blue's success at chess, other games have been mastered by machines such as IBM Watson winning Jeopardy, or DeepStack's poker triumph, dominating an imperfect information game requiring intuition. The victory of Google DeepMind's AlphaGo against the world master Lee Sedol in March 2016 marked a turning point, as the machine showed signs of what commentators referred to as creativity;
- Self-driving vehicles: equipped with sensors and analysing gigabytes of information each second, the new generation of automated vehicles combine different AI systems to drive themselves (Tesla or Waymo);
- Medical diagnosis: AI can help physicians establish or confirm a diagnosis (Human Dx);
- Killer robots: lethal autonomous weapons systems are able to select and engage targets with little or no human intervention;[2]
- 'Eureka machines' that support the creativity process to produce new inventions (help in the design of new objects and materials), optimise existing solutions, or find new solutions to problems without preconceived notions that might limit human creativity. These machines are also creating new recipes (Chef Watson) or can help mathematicians explore new areas of pure mathematics;
- Artistic machines, such as AI able to write stories or create artworks such as paintings (The Painting Fool) or musical compositions (Aiva).

AI systems are also used for customer services, as personal assistants (Siri or Sally), to make purchasing predictions and recommendations on online stores (Otto), or to ensure surveillance. For many of this tasks AI systems are expected to perform better than humans in the near future (mid-2020s).

Current limitations and the future of artificial intelligence

Narrow or weak artificial intelligence: a new form of intelligence

Data-driven AI is also referred to as 'narrow AI' or 'weak AI' because it creates machines that are only able to do one task very well: recognise cats (not cows); play go (not chess); invent a recipe (not compose music). Self-driving cars operate through the combination of various one-task AI systems that together are able to provide a three-dimensional map of the surroundings of the vehicle so that it can make appropriate decisions. Consequently, the skills that such machines can acquire to perform a task are not transferable to another task, which is an important aspect of human-like intelligence.

Data-driven AI can be seen as a new form of intelligence, different to what the human brain can perform. It allows machines to do tasks as well as humans, but much faster. By using statistical correlation from a huge quantity of data, machines are able to perform tasks that would require intelligence when performed by humans. However, machines do so using non-human methods, without replicating the performances of the human mind.

Artificial intelligence systems lack common sense and intentionality

It is still not possible for machines to understand what would come next in a series of images; to understand the broader context of a scene in a given image. This capacity to understand the 'rules of the world', a basic skill in humans at an early age, is referred to as having 'common sense'. Machines lack common sense, and therefore do not understand what a given object is, what it is used for, and where it is usually located, whereas even human children understands that the pan in the kitchen is used to cook.

As a lack of common sense could be linked to an absence of other human capacities, such as feelings or emotions, some researchers have taken this direction of inquiry. Others think that intelligence can only result if machines can sense the outside world with a body and learn how to adapt. Unsupervised learning may bring machines closer to acquiring common sense, but also to learning how to manage uncertainty. Unsupervised learning is increasingly used so that machines try to learn tasks without being taught: learning to play games by watching them for example.

Machines remain unable to share a goal with a human. They are programmed to carry out a task that might help humans reach a goal, but they are not able to share this goal or to understand it. An autonomous car is a tool to get from point A to point B, it cannot share a passenger's goal to take a route that offers better sightseeing.

Machines do not share intentionality with their operator: they are tools that are unable to collaborate, regardless of how intelligent they appear. This has considerable implications for interaction between humans and AI systems, in order to ensure that humans do not rely blindly on machines because they seem intelligent. Humans should be able to take over and correct the system when something unexpected occurs. Implications can be dramatic, as demonstrated by the death of the driver of an automated Tesla car in May 2016.

Artificial general intelligence: reaching 'strong artificial intelligence'

The original AI quest was to create machines that are able to display the same level of intelligence as humans: referred to as artificial general intelligence or 'strong AI'. Strong AI would mean that machines would perform different tasks, show common sense and share intentionality. They would display creativity and intuition, perhaps even emotions and consciousness. It might also be possible for these machines to set their own goals and values. They would display the full range of human capacities with the potential to do everything faster and better.

An extrapolation of such a situation is to expect that a human-like intelligent machine would be able to self-improve, leading to an exponential increase of their intelligent capacity. Such a superintelligence, outstripping human intelligence, would lead to a 'technological singularity', leaving humans in the hands of machines. However, there is no evidence that human intelligence is a tipping point beyond which such an increase is possible. Displaying a higher level of intelligence means more than thinking faster and stronger about issues. Moreover, many current complex issues would not be solved with an exponential improvement of computational capacities. Such a hypothetical exponential development would also be limited and constrained by physical laws.

Researchers in the field agree that, contrary to frequent announcements, reaching strong AI is decades, or farther, away. In the meantime, the inability to tackle the limitations of narrow AI might lead to a decrease in enthusiasm for AI applications and the onset of new AI winters.

Key issues in data-driven artificial intelligence

Quality and privacy of the data

In order to perform tasks, current AI systems feed on a very large quantity of data that is either provided to them or collected by them. Making sure that the collection, access, use and storage of data for AI applications does not threaten end-users' privacy is one of the challenges of data-driven AI.

Nevertheless, another issue relates to the quality of data. Existing bias in the data used to train AI systems might lead to unwanted bias in the decisions it takes. It has been observed that AI systems can display ethnic- or gender-based bias in their decision-making. Machine learning systems not only learn prejudices from data but they can also amplify bias and inequality in the decisions they make.

One solution is to ensure the quality of the data used, to avoid bias or to make sure that the program does not take any bias into account. This last option may be difficult, due to the opacity regarding how AI systems reach their conclusions.

Autonomous decision-making

The biggest issue in data-driven AI is to decide how to control these systems and whether they are given the possibility to make autonomous decisions. This question is even more sensitive considering that there is no way to understand how narrow AI systems reach their decisions. Whatever level of intelligence AI systems appear to display, these systems do not understand the actions they perform, nor the goal that they are asked to pursue for humans. They are tools, whose core business is to provide recommendations and probabilistic predictions. In that sense, some researchers talk about 'stupid AI'.

It is always up to a human to decide whether these decisions can be made automatically, without further control, or if human intervention is needed. This is of particular importance for all the life-changing decisions that can be automated, such as those regarding mortgages, recruitment, or health. It is also of paramount importance when discussing automated weapons.

A regulatory framework to control algorithms and their impact is therefore important, including the possibility of engaging independent auditors for algorithms (or even software watchdogs), or a regulator that can investigate AI automated decisions. To avoid the erosion of human privacy and autonomy, it is vital to resist the temptation to use AI systems before appropriate cultural and legal frameworks are adapted.

Impact of artificial intelligence on the job market

The recent progress in AI and the impact on the development of robotics raises fears regarding the evolution of the job market, as jobs seem increasingly at risk of being automated in the coming decades. AI researchers estimate the full automation of human work around 120 years from now. Job automation depends on the type of task and the competencies needed. Some medical professions seem to be harder to replace with AI systems. Telemarketers are often ranked as the profession most at risk of replacement in the short-term.

Nevertheless, AI's impact on the job market will depend on how much autonomy is given to AI systems to make decisions and how much control will be required for such systems, i.e. how the interactions between AI systems and workers will be managed. AI can be used as a tool to upskill workers or help them to complete some tasks. AI systems also require training and monitoring. Humans will be needed to fill the gaps or manage unexpected situations that AI systems would not be able to handle, creating new types of work.

Artificial intelligence systems and the rise in inequality

Beyond the impact on the job market, discussions are ongoing on the impact of AI systems on social inequality. Some forecast that the automation of certain jobs would reduce the possibility for social mobility and benefit the wealthiest, able to adapt better to the change. AI tools may reshape how wealth is created and alter the global balance of power, leading to further inequality. Others argue that AI holds great potential to redistribute wealth and that the wealth concentration scenario can be avoided. The issue of AI and inequality is currently discussed in international fora, such as the United Nations or the OECD.

New frameworks for the development of artificial intelligence

Framing the development of data-driven artificial intelligence in Europe

In order to address the issues posed by data-driven AI, the EU's legal frameworks regarding data, algorithms and robots have to be modified. In April 2016, the European Parliament and the Council of the European Union adopted the General Data Protection Regulation (GDPR) to frame the collection, use and storage of data in the EU. Another key aspect of this regulation is to allow citizens the opportunity 'not to be subject to a decision based solely on automated processing' (Article 22).

This regulation led to the adoption of guidelines on automated individual decision-making and profiling in October 2017. These guidelines require data controllers to 'find simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision without necessarily always attempting a complex explanation of the algorithms used or disclosure of the full algorithm'. This is likely to be a challenge for AI systems developers.

The European Parliament adopted a resolution on civil law rules on robotics in February 2017. The Parliament called on the European Commission to provide definitions of the category of smart robots. It considered that the Union legal framework on robotics and AI should be updated and complemented with guiding ethical principles. The Parliament highlighted a principle of transparency requesting that 'it should always be possible to supply the rationale behind any decision taken with the aid of an AI that can have a substantive impact on one or more persons' lives'.

The Parliament called for the designation of a European Agency for Robotics and Artificial Intelligence to provide support on these issues. It asked the Commission to submit a proposal for a legislative instrument on legal questions related to the development and use of robotics and AI and guidelines and codes of conduct regarding robotics and the ethical questions associated.

In its May 2017 communication on the midterm review of the implementation of the digital single market strategy, the Commission noted that it will consider the need to adapt current legislation to take into account the developments on robotics and AI. The Commission also declared its ambition for the EU to take the lead in the development of AI technologies, platforms and applications, reaffirming a position already expressed in a communication on the strategy for digitising European industry of April 2016. The Commission also adopted a proposal for a regulation on a framework for the free flow of non-personal data in the European Union in September 2017.

Preparing the frame for future artificial intelligence

In the long-term, the potential development of strong AI will imply deep modifications of legal and cultural frameworks. It is vital to design future intelligent machines wisely: if they are to set their own goals and values, these should coincide with human goals and values. Intelligent machines should also not be able to reach a given goal by making cold-blooded rational choices that could counter humans' interests. These objectives are expected to guide future developments in AI.

Main references

New Scientist, *Machines that think*, Hodder & Stoughton, October 2017.

Slowman S. & Fernbach P., *The knowledge illusion*, Riverhead Books, March 2017.

Endnotes

[1] AI systems are software able to perform given tasks in imitation of human intelligent behaviours. AI is a branch of computer science. AI should not be confused with robotics – which is an engineering science. Robotics deals with the construction and use of hardware systems (robots) managed by computer systems that can be AI systems.

[2] A campaign to stop killer robots was launched by an international coalition of NGOs on 13 November 2017.

Disclaimer and Copyright

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

Photo credits: © ktsdesign / Fotolia.

eprs@ep.europa.eu
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)