European Commission

JRC TECHNICAL REPORTS

AI Watch
Historical Evolution of Artificial Intelligence

Analysis of the three main paradigm shifts in AI

EUR 30221 EN

Joint Research Centre

Contact information
Blagoj DELIPETREV
European Commission, Joint Research Centre, TP262, Via Fermi, 21027 Ispra (VA), ITALY
blagoj.delipetrev@ec.europa.eu
Tel.: + 39 0332 786 352

AI-Watch functional mailbox
EC-AI-WATCH@ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

JRC120469

EUR 30221 EN

PDF ISBN 978-92-76-18940-4 ISSN 1831-9424 doi:10.2760/801580

Luxembourg: Publications Office of the European Union, 2020

How to cite this report: Blagoj DELIPETREV, Chrisa TSINARAKIi, Uroš KOSTIĆ. "*Historical Evolution of Artificial Intelligence*", EUR 30221EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-18940-4, doi:10.2760/801580, JRC120469

Contents

Foreword

This report is published in the context of AI Watch, the European Commission knowledge service to monitor the development, uptake and impact of Artificial Intelligence (AI) for Europe, launched in December 2018.

AI has become an area of strategic importance with potential to be a key driver of economic development. AI also has a wide range of potential social implications. As part of its Digital Single Market Strategy, the European Commission put forward in April 2018 a European strategy on AI in its Communication "Artificial Intelligence for Europe" COM (2018)237. The aims of the European AI strategy announced in the communication are:

- To boost the EU's technological and industrial capacity and AI uptake across the economy, both by the private and public sectors
- To prepare for socio-economic changes brought about by AI
- To ensure an appropriate ethical and legal framework.

Subsequently, in December 2018, the European Commission and the Member States published a “Coordinated Plan on Artificial Intelligence”, COM (2018)795, on the development of AI in the EU. The Coordinated Plan mentions the role of AI Watch to monitor its implementation.

AI Watch monitors European Union’s industrial, technological and research capacity in AI; AI-related policy initiatives in the Member States; uptake and technical developments of AI; and AI impact. AI Watch has a European focus within the global landscape. In the context of AI Watch, the Commission works in coordination with Member States. AI Watch results and analyses are published on the AI Watch Portal (https://ec.europa.eu/knowledge4policy/ai-watch_en).

From AI Watch in-depth analyses, we will be able to understand better European Union’s areas of strength and areas where investment is needed. AI Watch will provide an independent assessment of the impacts and benefits of AI on growth, jobs, education, and society.

This report addresses the following objectives of AI WATCH: Analysis of the evolution of AI technologies. As part of this objective this report particularly aims to analyse the historical evolution of AI and the similarities and differences among the different phases of this evolution.

This is a JRC technical report within the context of our Unit activities on AI.

Acknowledgements

We are grateful to Professor Barry O'Sullivan from the Insight Centre for Data Analytics, School of Computer Science & IT, University College Cork and Professor José Hernández-Orallo Universitat Politècnica de València, whose insightful comments assisted us in further improving the report.

We are thankful to our JRC colleagues: Emilia Gómez, Massimo Craglia, Paul Desruelle, Giuditta De Prato, Montserrat López Cobo, Marius Miron and Alessandro Annoni, for their valuable comments and suggestions.

Authors

Blagoj Delipetrev, Chrisa Tsinaraki, Uroš Kostić.

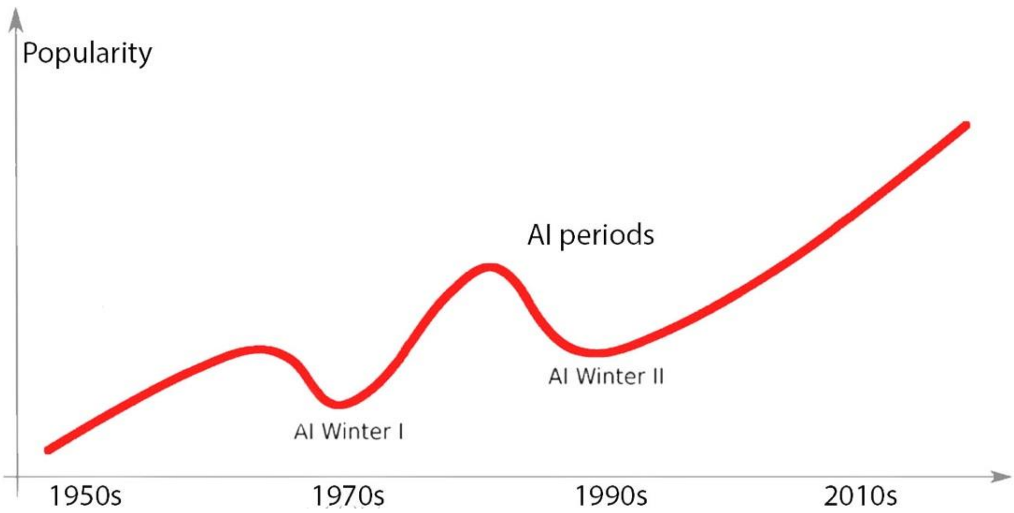
Executive summary

Today, AI has become a mature technology and an increasingly important part of the modern life fabric, but its future is uncertain, with chances of both a higher growth and a decline. Reflecting on the history of the domain may provide us with insights on its future. Starting with fundamental definitions and building on the historical context, this report summarizes the evolution of AI, it presents the periods of AI development and describes the current rise of interest in AI. It concludes with a comparison of the current and past AI development contexts, trying to get insights on the uncertain future of AI.

This report presents:

1. AI fundamentals, including several AI definitions, main approaches and methods.
2. The AI development periods ("seasons" - i.e. winters for the decline and springs for the growth), shown in Figure 1, including:
   - The spring of the 1950s, with the establishment of the foundation of most of the AI algorithms, followed by the winter of the 1970s.
   - The spring of the 1970s, with the paradigm shift in symbolic algorithms and building of expert systems, followed by the winter of the 1990s.
   - The development of machine learning in the 1990s, which led to the development of deep learning and the spring of the 2010s.

Figure 1: Illustrative visualization of AI periods



Source: JRC

3. An analysis of the AI paradigm shifts accompanied by bold predictions, followed by massive investments, and major issues. Some of the major issues in the first two springs have led to the two AI winters. The analysis is more focused on the current machine learning and deep learning period.
4. A comparison of the current and past AI development contexts to get insights on the uncertain future of AI. The context parameters examined include, among others, the main AI investors, the entities performing basic research, the paradigm shifts, the computing capacity, the availability of relevant regulation, etc.

The future of AI is uncertain, with chances of another AI winter or of an even greater AI summer. Given this uncertainty, it is particularly important to have in place an unbiased initiative such as AI Watch to monitor the evolution and assess its impacts over the coming years which one way or the other will see very significant changes in our digitally transformed society.

Abstract

Artificial intelligence (AI) can have a major impact on the way modern societies respond to the hard challenges they face. Properly harnessed, AI can create a more fair, healthy, and inclusive society. Today, AI has become a mature technology and an increasingly important part of the modern life fabric. AI is already deployed in different application domains, e.g. recommendation systems, spam filters, image recognition, voice recognition, virtual assistants, etc.

It spans across many sectors, from medicine to transportation, and across decades, since the term was introduced in the 1950s. The approaches also evolved, from the foundational AI algorithms of the 1950s, to the paradigm shift in symbolic algorithms and expert system development in the 1970s, the introduction of machine learning in the 1990s and the deep learning algorithms of the 2010s.

Starting with the fundamental definitions and building on the historical context, this report summarizes the evolution of AI, it introduces the “seasons” of AI development (i.e. winters for the decline and springs for the growth), describes the current rise of interest in AI, and concludes with the uncertainty on the future of AI, with chances of another AI winter or of an even greater AI spring.

1 AI definitions and main classifications

AI is defined as machine intelligence or intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. The term AI is often used to describe machines that mimic human cognitive functions such as learning, understanding, reasoning or problem-solving (Russell and Norvig 2016). AI has two main dimensions, as shown in Table 2 below. The AI definitions on top of the table relate to processes and reasoning, whereas the ones at the bottom side address behaviour. The definitions on the left side of the table measure success in terms of fidelity to human performance, whereas the ones on the right-side measure against an ideal concept of intelligence and rationality.

Table 1: Some definitions of AI, organized in four categories

Systems that think like a human	Systems that think rationally
"The exciting new effort to make computers think … machines with minds, in the full and literal sense" (Haugeland 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem-solving, learning …" (Bellman, 1978)	"The study of mental faculties through the use of computational models" (Chamiak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston 1992)
A system that acts like humans	**A system that acts rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Ritch and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole at al. 1998) "AI … is concerned with intelligent behaviour in artefacts." (Nilsson, 1998)

Source: (Russell and Norvig 2016)

The High-Level Expert Group on Artificial Intelligence (HLEG 2019) produced an AI definition that has also been adopted by AI Watch (Samoili, Lopez Cobo, et al. 2020):

"*Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.*

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors, and actuators, as well as the integration of all other techniques into cyber-physical systems)."

Artificial Narrow Intelligence (ANI), often referred to as "Weak" AI is the type of AI that mostly exists today. ANI systems can perform one or a few specific tasks and operate within a predefined environment, e.g., those exploited by personal assistants Siri, Alexa, language translations, recommendation systems, image recognition systems, face identification, etc.

ANI can process data at lightning speed and boost the overall productivity and efficiency in many practical applications, e.g., translate between 100+ languages simultaneously, identify faces and objects in billions of images with high accuracy, assist users in many data-driven decisions in a quicker way. ANI can perform routine, repetitive, and mundane tasks that humans would prefer to avoid.

While ANI is superior in specialized domains, it is incapable of generalization, i.e. to re-use learned knowledge across domains, e.g., the ANI capable of image recognition cannot transfer its knowledge in the domain of speech recognition. The generalization problem is still an open question (Hernández-Orallo 2017).

Artificial General Intelligence (AGI) or "Strong" AI refers to machines that exhibit human intelligence. In other words, AGI aims to perform any intellectual task that a human being can. AGI is often illustrated in science fiction movies with situations where humans interact with machines that are conscious, sentient, and driven by emotion and self-awareness. At this moment, there is nothing like an AGI.

Artificial Superintelligence (ASI) is defined as "*any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*" (Bostrom 2016). ASI is supposed to surpass human intelligence in all aspects — such as creativity, general wisdom, and problem-solving. ASI is supposed to be capable of exhibiting intelligence that we have not seen in the brightest thinkers amongst us. Many thinkers are worried about ASI. At this moment, ASI belongs to science fiction.

If we ever succeed in creating an AI that is capable of generalizing, understanding causality, making a model of the world, it is highly likely that it will be closer to ASI than AGI. AI excels in numerical calculations, and there is no logical explanation as to why AI would downgrade its abilities to simulate humans. AI's quest ultimately leads to ASI.

ML is an AI subfield. Machine learning (ML) is the scientific study of algorithms that computer systems that learn through experience. ML algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. ML can be divided into these categories:

- Supervised learning algorithms map input to output values based on labelled examples of input-output pairs. For example, we want to predict whether the image contains a cat, then the algorithm is shown a lot of labelled images with cats and without cats. Afterwards, the ML algorithm learns to recognize cats in unseen images. Supervised learning needs considerable amounts of labelled data, which is often done by humans.
- Unsupervised learning algorithms help finding previously unknown patterns in datasets without pre-existing labels. The objective is to discover the underlying data structure, for example, by grouping similar items to form "clusters." Unsupervised learning does not require labelled data, but instead tries to learn by itself.
- Semi-Supervised learning algorithms can be considered a category between supervised and unsupervised learning, where the data contains both labelled and unlabelled data.
- Reinforcement learning (RL) explores how agents take actions in an environment in order to maximize a reward. An example is when the RL agent plays Go against itself, learns the game, and acquires above human intelligence in Go.

2 AI Foundations (1950s – 1970s)

In 1950, Alan Turing published the milestone paper "Computing machinery and intelligence" (Turing 1950), considering the fundamental question "Can machines think?" Turing proposed an imitation game, known as the Turing test afterwards, where if a machine could carry on a conversation indistinguishable from a conversation with a human being, then it is reasonable to say that the machine is intelligent. The Turing test was the first experiment proposed to measure machine intelligence.

The first "AI period" began with the Dartmouth conference in 1956, where AI got its name and mission. McCarthy coined the term "artificial intelligence," which became the name of the scientific field. The primary conference assertion was, "*Every aspect of any other feature of learning or intelligence should be accurately described so that the machine can simulate it*" (Russell and Norvig 2016). Among the conference attendees were Ray Solomonoff, Oliver Selfridge, Trenchard More, Arthur Samuel, Herbert A. Simon, and Allen Newell, all of whom became key figures in the Ai field

People were excited because for the first-time computers were solving problems like humans and seemed intelligent. The wider AI research community shared an initial optimism making bold claims and boosting popularity. Examples, where AI solved problems included algebraic application problems, language translation, geometric theorem proving, etc. A list of several important AI breakthroughs of that period are:

1952	Checkers was the first program to demonstrate that computers can learn and not just perform what they are programmed to do. Checkers attracted media attention and learned to play at a level high enough to challenge a decent amateur human player (Samuel 1960).
1955	The Logic Theorist had proven 38 theorems from Principia Mathematica and introduced critical concepts in artificial intelligence, like heuristics, list processing, 'reasoning as search,' etc. (Newell et al. 1962).
1957	Inspired by the human brain, Rosenblatt discovered the perceptron. The perceptron was predicted to be "the embryo of an electronic computer that will be able to walk, talk, see, write, reproduce itself and be conscious of its existence." The perceptron was the birth of connectionism, the foundation of Neural Networks (NN) and Deep Learning (Rosenblatt 1961).
1961	Machine Educable Nougats And Crosses Engine (MENACE) was one of the first programs capable of learning to play a perfect game of Tic-Tac-Toe (Michie 1963).
1965	ELIZA was a natural language processing system that imitated a doctor. ELIZA responded to questions like a psychotherapist. Some users believed they are interacting with another human being until it reached its limitations, and the conversation became nonsense (Weizenbaum 1966).
1969	Shakey the Robot was the first general-purpose mobile robot capable of reasoning its actions. This project integrated research in robotics with computer vision and natural language processing, thus being the first project that combined logical reasoning and physical action (Bertram 1972).
1969	The book "Perceptrons" highlighted unrecognized limits of the feed-forward, two-layered perceptron structure. The authors' pessimistic predictions made a fundamental shift in the AI research direction to symbolic and disregarding connectionism. "Perceptrons" marks the beginning of the AI winter of the 1970s (Minsky and Seymour 1969).

2.1 Bold predictions

Leading researchers' bold predictions:

Herbert Simon, 1957:

"*It is not my aim to surprise or shock you, but the simplest way I can summarize is to say that there are now in the world machines that think, that learn, and that create. Moreover, their ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be coextensive with the range to which the mind has been applied.*"

Marvin Minsky, 1970:

"*In three to eight years, we will get a machine with average human intelligence.*"

2.2 Investments

Intrigued by the expectations and the optimism of the leading scientists, the US government massively funded AI research. J.C.R. Licklider, the director of ARPA, claimed "*fund people, not projects*." He believed that ARPA should allow researchers to follow their direction of interest. The founding principle resulted in an unconstrained research atmosphere with a significant money flow without any hard products and deliverables. ARPA later became The Defence Advanced Research Projects Agency (DARPA). DARPA invested millions in supporting the academic AI research centres like the AI lab at the Massachusetts Institute of Technology (MIT) with $2.2 million, the Stanford University AI project established by John McCarthy in 1963, the Carnegie Mellon University, as well as commercial research labs such as SRI International.

2.3 Issues

The first AI winter started in the 1970s, due to unfulfilled promises, vast expectations, and financial difficulties. At the same time, AI encountered impossible-to-overcome technological barriers, mainly on the limitations of computing power, memory, and processing speed. In the mid-1960s, researchers discovered a distinction between polynomial and exponential growth in problem complexity. The exponential growth in complexity prohibits solving moderately large instance problems at any reasonable time. This led to the most important concept of complexity, NP-completeness, and its most fundamental question, whether P = NP, where "P" is a general class of questions for which some algorithm can provide an answer in polynomial time. Many combinatorial and logical problems are NP-complete requiring exponential processing time and unfeasible-to-be-effective systems.

The first AI Winter has been marked by a dramatic decrease in the AI activities in both industry and academia. The AI shortcomings were explained in two reports: a) the Automatic Language Processing Advisory Committee (ALPAC) report by the US Government (ALPAC 1966) and b) the Lighthill report (Lighthill 1973) by the British government.

2.3.1 The ALPAC report

During the cold war, automatic machine translation from Russian to English was among the main US interests. The machine translation Georgetown-IBM experiment was the New York Times front page in an article entitled, "*Russian is turned into English by a fast-electronic translator*." Reports appeared in newspapers and magazines, citing the experiment leader's Leon Dostert prediction that soon meaningful machine translation will be accomplished.

These predictions established massive funding in machine translation by the US agencies since 1956. Progress was slow, and the ALPAC committee was assigned to report on the reasons. The ALPAC report focused on the economic return of machine translation without considering its scientific value. The report concluded that machine translation did not provide adequate improvements in quality, speed, and cost; on the contrary, machine translation was of poor quality. It noted that enough translators were available (4,000 were currently under contract, but only 300 of them on average worked each month). Government officials were disappointed, emphasizing the little benefits gained from the massive investments.

2.3.2 The Lighthill report

The British Science Research Council published a report titled "*Artificial Intelligence: A General Survey*" by Prof. Sir James Lighthill from the University of Cambridge in 1973. The Lighthill report was pessimistic since "*in no part of the AI field have discoveries made so far produced the major impact that was then promised.*"

The report was reflecting on AI disappointment in both the public and the scientific community. The report highlighted that the conventional engineering approach with radio waves was performing better than AI methods of automatic landing systems for airplanes. AI experiments worked in the labs and small domains, but were inadequate in large-scale real-world problems, often because of the "Combinatorial Explosion."

3 Symbolic AI (1970s – 1990s)

In the 1980s, the AI paradigm shifted to symbolic AI and the so-called "expert systems" or "knowledge-based systems." The underlying idea was to get human expert knowledge in a computer form and spread it as a program to many Personal Computers (PCs). Expert systems had two components: 1. the knowledge base – a collection of facts, rules, and relationships on a specific domain; and 2. the inference engine that described how to manipulate and combine these symbols. The facts and rules had explicit representation and were modifiable. Lisp and Prolog were the main symbolic programming languages.

Companies started producing expert systems in the 1990s. These companies offered software packages called "inference engines" and related knowledge services to customers. However, these tools and frameworks lacked sufficient expressive power to capture the breadth of the expert knowledge and behaviour required to achieve satisfactory performance. Several important AI breakthroughs of that period are:

1970	MYCIN was an expert system specialized in the diagnosis of blood diseases and prescription of drugs. It adopted a calculus of uncertainty that seemed to fit well with the doctors' assessment on the diagnosis (Shortliffe et al. 1975).
1968-1970	SHRDLU was a natural language computer program that allowed the user to carry on a conversation with the computer, name collections, and query the state of a simplified "blocks world" (Winograd 1971).
1972	Prolog is a symbolic programming language developed and implemented in Marseille, France, by Alain Colmerauer with Philippe Roussel.
1979	The Stanford Cart managed to cross a room on its own while avoiding obstacles (Moravec 1983).
1982	"Hopfield net" was a form of neural network that learned and processed information in a new way (Hopfield 1982). The "Hopfield net" and "backpropagation" (Rumelhart et al. 1985) revived the AI field of connectionism.
1983	ID3 is an algorithm that generates a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm used in machine learning and natural language processing (Quinlan 1986).

3.1 Bold predictions

Hans Moravec, 1977:

"*Suppose my projections are correct, and the hardware requirements for human equivalence are available in 10 years for about the current price of a medium-large computer. Suppose further that software development keeps pace (and it should be increasingly easy, because big computers are great programming aids), and machines able to think as well as humans begin to appear in 10 years.*"

Marvin Minsky, 1997:

"*I'm still a realist: If we work really hard - and smart - we can have something like a HAL in between four and four hundred years. I suppose if we're lucky, then, we can make it by 2001!*"

3.2 Investments

The expert system called XCON saved $40 million a year to the Digital Equipment Corporation in 1980-1986. The leading hardware companies were Symbolics, Lisp Machines, and software companies IntelliCorp and Aion. By 1985, the US investments in AI expert systems reached over $1 billion. The UK started the £350 million Alvey project.

In 1981, the Japanese Ministry of Economy, Trade, and Industry allocated $850 million to support the fifth-generation computer project. The goal was to create machines that could talk to people, translate languages, interpret images, and reason like humans.

The European Strategic Program on Research in Information Technology (ESPRIT) was a series of integrated programmes of information technology research and development projects and industrial technology transfer measures. It was a European Union initiative managed by the Directorate General for Industry (DG III) of the European Commission. ESPRIT had a substantial budget comparable to the USA and Japan.

3.3 Issues

After the 1990s, the term “expert system” dropped from the IT lexicon, and it was referred to as the second AI winter. One of the main problems in expert systems was knowledge acquisition. Knowledge acquisition captures expert knowledge and represents it in a symbolic language. Obtaining domain expert time and expertise was difficult since they are in constant need by their organizations. Therefore, expert system research focused on tools for knowledge acquisition, to help automate the process of designing, debugging, and maintaining rules defined by the experts.

The symbolic programming languages were Lisp and Prolog, and the hardware platforms were the Lisp machines and PCs. The expert system development environment could not match the compiled language efficiency (such as C). The PC performance of Apple and IBM continued to increase, and by 1987, they had surpassed the expensive specialized Lisp machines produced by Symbolics and other manufacturers, collapsing a $500 million industry.

However, while funding dried up, and researchers avoided the term AI, many of them continued working under other discipline-specific names: cognitive systems, intelligent systems, knowledge representation, and reasoning. The Business Rules Management System (BRMS) is a legacy of the symbolic era. BRMSs are still used across many industries.

4 Machine learning and deep learning (1990s – 2020s)

In the 1990s–2010s, AI had addressed complex problems, providing solutions that were found to be useful in different application domains including data mining, industrial robotics, logistics, business intelligence, banking software, medical diagnosis, recommendation systems, and search engines. AI researchers began to develop and use more sophisticated mathematical tools. There was a widespread realization that many AI problems have already been worked on by researchers in fields like mathematics, economics, or operations research. The shared mathematical language allowed a higher level of collaboration with established fields and made AI a more rigorous scientific discipline.

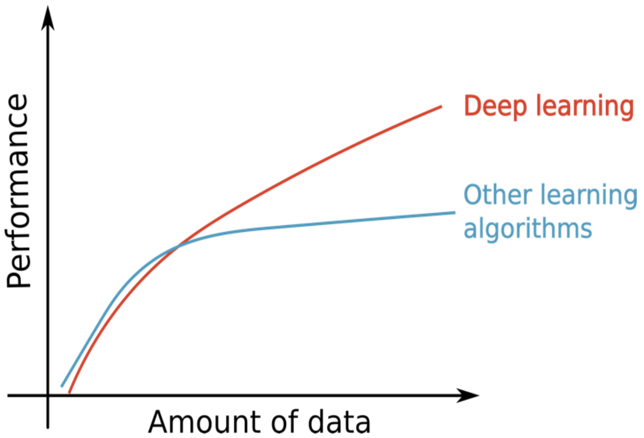
Many AI researchers in the 1990s deliberately called their work by other names, such as informatics, knowledge-based systems, cognitive systems, optimization algorithms, or computational intelligence. The new names helped to procure funding. The failed promises of the AI Winter continued to haunt AI research in the commercial world.

In 2006, Fei-Fei Li, a computer science professor at Stanford University, contributed to a paradigm change. The established wisdom of the researchers at that time was that *"If you can't even do one image well, why try with thousands or tens of thousands of images?"* Her hypothesis was that AI's main limitation is the data quantity to reflect the real-world scenarios and that more data will produce better models. In 2009, ImageNet was published containing 3.2 million labelled images, separated into 5,247 categories, sorted into 12 subtrees like "mammal", "vehicle", "furniture", etc.' (Deng et al. 2009).

The annual competition ImageNet Large Scale Visual Recognition Challenge (ILSVRC) saw dramatic progress in the last decade. The average ILSVRC classification error rate was around 25% in 2011. In 2012, a deep convolutional neural net called AlexNet achieved a 16% classification error rate (Krizhevsky et al. 2012), and in the next couple of years, error rates fell to a few percent. These breakthroughs made the AI paradigm shift to deep learning (DL).

DL performance improves with the amount of data compared to other ML algorithms, as shown in Figure 2.

Figure 2: Deep Learning performance considering the amount of data

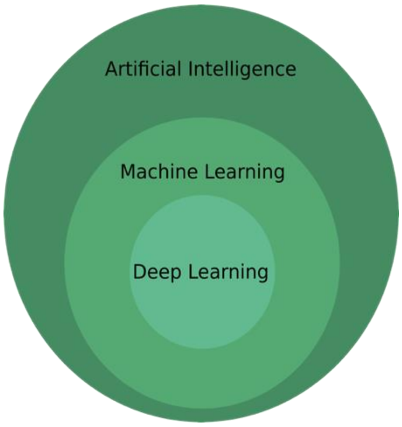


Source: JRC

DL is a subfield of ML and AI, as shown in Figure 3. DL introduced a multi-layer neural network architecture that learns data representations with levels of abstraction (LeCun et al. 2015). DL neural network architectures include deep neural networks, deep belief networks, recurrent neural networks, and convolutional neural networks (CNN). The terms AI or ML often replace DL, especially in the news and media.

The classical programming paradigm consists of defining explicit instructions in different programming languages such as Java or C. The developer specifies the rules, which, combined with the data, produce the answers. In the ML paradigm, inputs include data and answers, and ML produces rules from inputs, as shown in Figure 4. The ML system is trained, rather than explicitly programmed. ML programming was an established paradigm, but got a new meaning when rules are saved as weights in the neural network. Typical DL networks can have millions or billions of weight parameters.

Figure 3: Artificial intelligence, machine learning, and deep learning

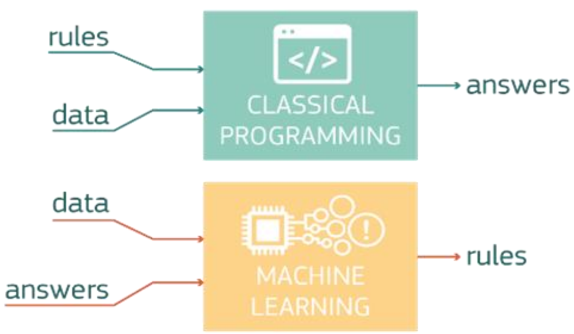


Source: JRC

For example, let us explain how DL and other classical algorithms would solve the problem of recognizing cats in an image. The classical algorithms would process the image using different methods to identify the two eyes, nose, legs, etc. and many other parts of a cat body. These methods required complex and long programs compared to DL, where the process of image recognition is almost completely automatic after selecting the DL architectural model (which, in this case, would probably be convolutional neural networks). DL automatically learns to recognize cats after observing many examples of images with and without cats.

Transfer learning is a technique whereby a neural network model is first trained on a large general problem, e.g., recognizing objects in an image. These pre-trained neural network models are then fine-tuned on our problem-specific image dataset. For example, pre-trained image recognition models such as VGG, InceptionV3 and ResNet5 can be fine-tuned for many image recognition problems, such as cat recognition. Transfer learning significantly speeds up the DL training stage, saves computational resources, and produces quality models.

Figure 4: Machine learning programming paradigm



Source: (Craglia et al. 2018)

The AI evolution is accelerated by the openness of code, frameworks, datasets, scientific publications, and overall knowledge sharing. Most AI algorithms are shared as open-source code that resides in GitHub, GitLab, or other code repositories. DL frameworks, e.g., TensorFlow, PyTorch, Theano, etc. are open source and supported by the largest IT companies such as Google or Facebook. Many of the latest AI research publications are freely available on arXiv[1], blogs, etc. ML competitions became preferred platforms, e.g., Kaggle, where users can share models, acquire datasets, collaborate, and compete in ML challenges. The popularity of massive open online course (MOOC) AI courses attracts millions of student’s worldwide, spreading AI knowledge and education, e.g.,

1 https://arxiv.org

MOOC courses offered by Coursera[2] since 2012 have had more than a million enrolments. These factors contributed to the broader and faster AI proliferation and adoption.

Some of the most important AI breakthroughs in this period are:

1989	Reading handwritten digits using convolutional neural networks. The system processed about 10-20% of handwritten cashed checks and zip codes in the United States between the late 90s and early 2000s (LeCun et al. 1989).
1989	"*Learning from Delayed Rewards*" introduced the concept of Q-learning, which greatly improves the practicality and feasibility of reinforcement learning. The Q-learning algorithm learned optimal control without modelling the transition probabilities or expected rewards of the Markov Decision Process (Watkins 1989).
1993	German computer scientist Schmidhuber solved a "very deep learning" task in 1993 that required over 1,000 layers in the recurrent neural network (Schmidhuber 1993).
1995	Support vector machines were applied to text categorization, handwritten character recognition, and image classification (Cortes and Vapnik 1995).
1995	"*No Hands Across America*[3]." A semi-autonomous car drove 4,501 km coast-to-coast across the United States with computer-controlled steering. A human driver-controlled throttle and brakes.
1996	IBM Deep Blue won the game of chess against Garry Kasparov, the best human player in the world. Deep Blue did not use DL or any of the latest AI techniques. The system had "learned" all possible played games of chess and could evaluate the present and suggest the next move.
1997	Long short-term memory (LSTM) architecture improved both the efficiency and practicality of RNN by eliminating the long-term dependency problem (Hochreiter and Schmidhuber 1997).
1998	Gradient-Based Learning was improved by combining the stochastic gradient descent algorithm with the backpropagation algorithm (LeCun et al. 1998).
2002	TD-Gammon achieved performance as the best human players in the game of Backgammon. TD-Gammon was a major achievement of combining neural nets and Reinforcement Learning (RL) with the self-play method (Tesauro 2002).
2005	The Stanford robot won the DARPA Grand Challenge. The Stanford robot drove autonomously for 131 miles along an unrehearsed desert track (Thrun et al. 2006).
2009	ImageNet dataset was launched with 3.2 million labelled images available to everyone (Deng et al. 2009).
2011	IBM Watson won a game of Jeopardy against Ken Jennings and Brad Rutter. Ken Jennings and Brad Rutter were among the most successful contestants on the Jeopardy show. Watson is a question-answering computing system with advanced natural language processing, information retrieval, knowledge representation, automated reasoning, and question answering. It was a truly remarkable AI system combining many state-of-the-art components in speech recognition, voice synthesis, and information retrieval, among others (Ferrucci 2012).
2012	AlexNet won the ImageNet competition, often considered as the inflection point of DL (Krizhevsky et al. 2012).

2 https://www.edsurge.com/news/2018-06-07-andrew-ng-is-probably-teaching-more-students-than-anyone-else-on-the-planet-without-a-university-involved

3 https://www.cs.cmu.edu/~tjochem/nhaa/general_info.html

2012	The Cat experiment has learned to identify and recognize cats from 10,000,000 unlabelled images randomly taken from YouTube. The program performed nearly 70% better than previous attempts of unsupervised learning (Le 2013).
2014	Generative Adversarial Networks (GANs) are deep neural net architectures composed of two nets, pitting one against the other (thus the “adversarial”). GANs can learn to mimic any distribution of data and can generate content like any domain: images, music, speech, etc. (Goodfellow et al. 2014).
2015	DeepRL has mastered a diverse range of Atari 2600 games to superhuman level with only the raw pixels and scores as inputs. Atari games became standard benchmarks for AI algorithms. AI has excelled human players in most of them (Mnih et al. 2015).
2016	AlphaGo defeated Lee Sedol, the world’s number one Go player (Silver et al. 2016). Because of its complexity, the game of Go was considered out of AI reach for at least another decade. AlphaGo was upgraded a year later into a generalized and more powerful algorithm AlphaZero (Silver et al. 2017). AlphaZero learned how to play master-level chess in just four hours and defeated Stockfish (the top AI chess player) in a 100-game match — without losing a single game.
2017	The Transformer is a novel DL architecture based on a self-attention mechanism. Transformers are core algorithms in language modelling, machine translation, and question answering (Vaswani et al. 2017).
2017	Libratus has decisively beaten four leading human professionals in the two-player variant of poker called heads-up no-limit Texas hold'em (HUNL). Over nearly three weeks, Libratus played 120,000 hands of HUNL against the human professionals, using a three-pronged approach that included precomputing an overall strategy, adapting the strategy to actual gameplay, and learning from its opponent. Poker is an imperfect information game (Brown and Sandholm, 2018).
2018	OpenAI Five has defeated an amateur human team at Dota 2, exceeding human intelligence in a complex video game. Relative to previous successes like Chess or Go, complex video games capture the messiness and continuous nature of the real world (Pachocki et a. 2018).
2019	AlphaStar has defeated a top professional player in StarCraft II. AlphaStar decisively beat Team Liquid’s Grzegorz "MaNa" Komincz, one of the world’s strongest professional StarCraft players with 5-0. The matches took place under professional match conditions on a competitive ladder map and without any game restrictions (Vinyals et al. 2019).
2019	OpenAI has trained Dactyl[4], a human-like robot hand to manipulate physical objects with unprecedented dexterity. Dactyl is trained entirely in simulation and transfers its knowledge to reality, adapting to real-world physics. Dactyl learns from scratch using the same general-purpose RL algorithm.
2019	GPT-2[5] is a large-scale unsupervised language model that generates coherent paragraphs of text, achieves state-of-the-art performance in many language modelling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

AI methods have been exploited in various domains, including computer vision, speech recognition, natural language understanding, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection, games, etc. The AI field is expanding in many domains, making it difficult and complicated to track its evolution. Papers with code[6] is tracking the AI progress providing a free and open

4 https://openai.com/blog/learning-dexterity/

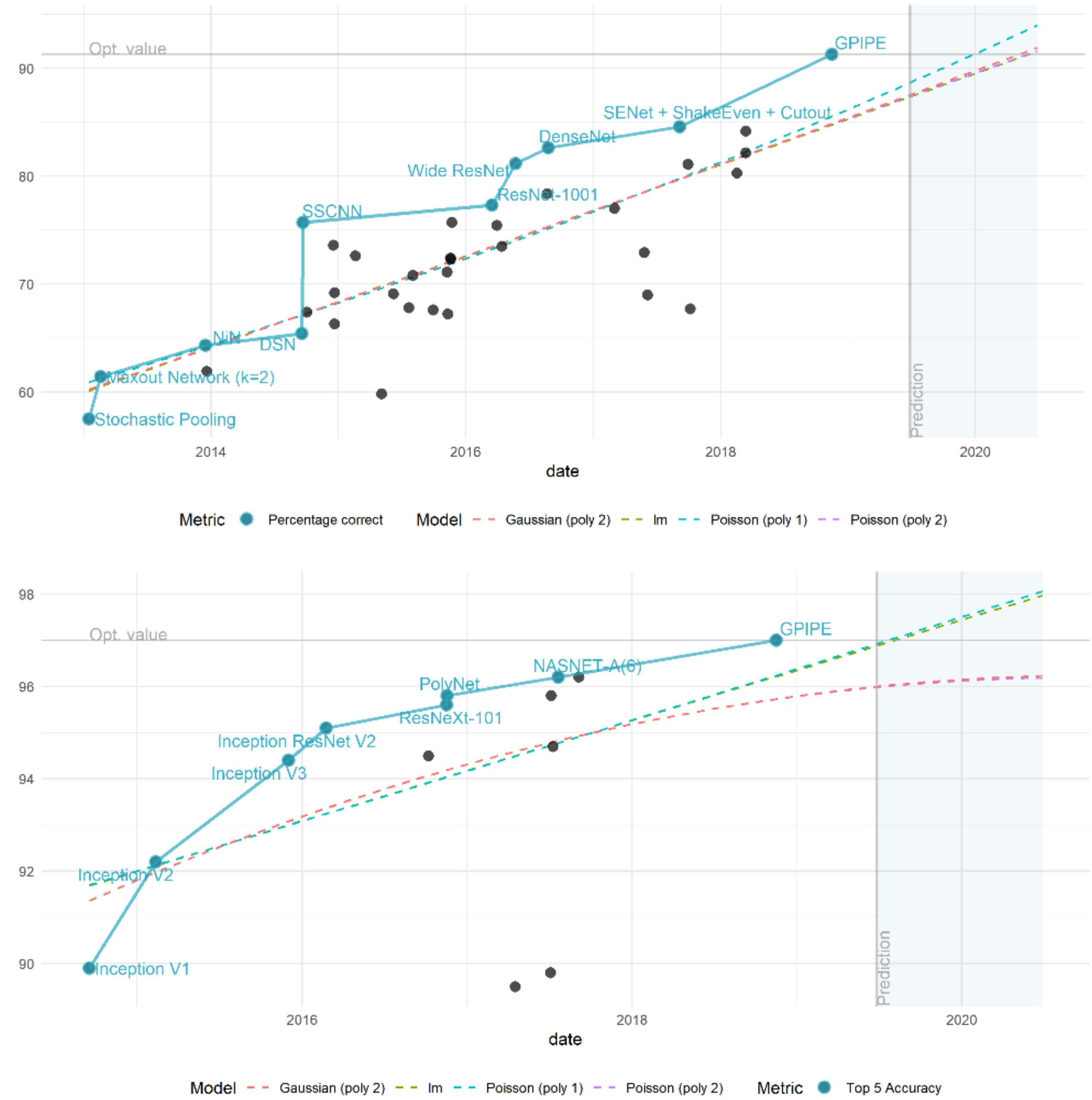
5 https://openai.com/blog/better-language-models/

6 https://paperswithcode.com/

repository of papers, code, and evaluation tables. The AI Collaboratory[7] aims to develop a collaborative initiative for the analysis, evaluation, comparison, and classification of AI algorithms. AI Collaboratory has become part of AI WATCH, European Commission's service for AI technology monitoring, and analysis.

We provide here some figures from the AI Collaboratory. We can observe the DL advancement in image recognition on the ImageNet dataset as illustrated in the top image in Figure 5. The y-axis shows model accuracy. Another historically famous dataset is CIFAR-100, with 80 million of tiny images in 100 classes containing 600 images (Figure 5, bottom image), showing a similar rate of progress in AI.

Figure 5: State-of-the-art table for Image Classification on ImageNet (top) and CIFAR 100 (bottom). Coloured dashed lines model the present and future (potential) trends using different linear and polynomial smoothing methods



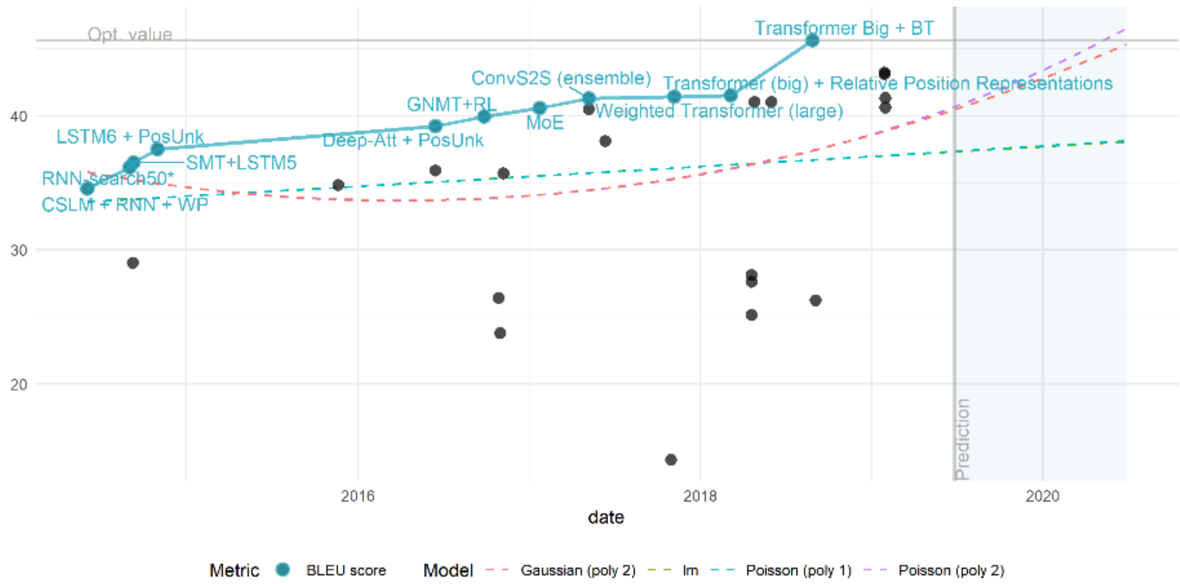
Source: http://dmip.webs.upv.es/AICollaboratory/

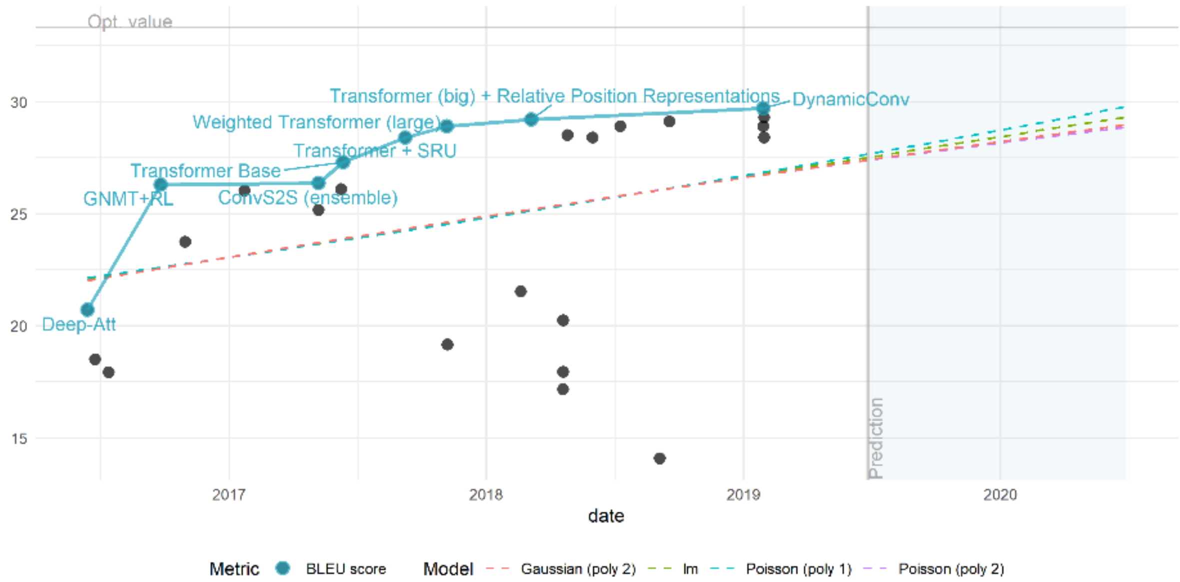
Machine translation performance is advancing with the continuous improvement of the Bilingual Evaluation Understudy Score (BLEU). BLEU is a metric for evaluating a generated sentence compared to a reference

7 http://dmip.webs.upv.es/AICollaboratory

sentence. The SOTA table for Machine Translation on WMT2014 for English to French (top) and English to German are shown in Figure 6. The DL language models made major advancements in 2018 with the introduction of transformers implemented in BERT and GPT-2 (Radford et al. 2019).

Figure 6: State-of-the-art table for Machine Translation on WMT2014 for English to French (top) and English to German (bottom). Coloured dashed lines model the present and future (potential) trends using different linear and polynomial smoothing methods





Source: http://dmip.webs.upv.es/AICollaboratory/

AI has transcended human intelligence in the games of Checkers 1952, Chess 1996, Jeopardy 2011, Go 2016, No-limit Poker 2017, Dota-2 2018. Checkers, Chess and Go are perfect information games as each player can observe all the pieces on the board. Card games where each player's cards are hidden from other players such as No-limit poker are examples of games with imperfect information. Dota 2 and StarCraft are complex and dynamic environments of multiplayer games. AI has succeeded to win human opponents in all of them.

With more board configurations than atoms in the observable universe, the ancient Chinese game of Go has long been considered a grand challenge for AI. On March 9, 2016, in South Korea, the top world player Lee Sedol collided with AlphaGo AI for an extraordinary best-of-five-game competition. Hundreds of millions of people watched as a legendary Go master took on an unproven AI challenger for the first time in history. The five games were astonishing, especially the two exceptional moves 1) 37 in game two by AlphaGo and 2) 78 in game four by Lee Sedol. Both moves at first seemed unreasonable but were decisive to win the game. In the end, AlphaGo won the series 4:1. AlphaGo learned by self-play, instead of studying the games based on human players. Some consider this competition as a turning point in China's AI strategy[8].

In 2018, Yoshua Bengio, Geoffrey Hinton, and Yann LeCun received the Turing Award for conceptual and engineering breakthroughs that have made DL a critical component of computing. They developed theoretical foundations for the field, identified surprising phenomena through experiments, and contributed engineering advances that demonstrated the practical advantages of deep neural networks.

4.1 Bold predictions

The statements below from scientists, researchers, investors, politicians, and historians, exhibit a wide variety of opinions and ideas, as well as predicted risks and gains of AI in the future.

Stephen Hawking, 2014:

"*The development of full artificial intelligence could spell the end of the human race… It would take off on its own, and re-design itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded.*"

Larry Page, 2006:

"*Artificial intelligence would be the ultimate version of Google. The ultimate search engine would understand everything on the web. It would understand exactly what you wanted, and it would give you the right thing. We're nowhere near doing that now. However, we can get incrementally closer to that, and that is basically what we work on.*"

Elon Musk, 2014:

"*I'm increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don't do something very foolish. I mean with artificial intelligence we're summoning the demon.*"

Ray Kurzweil, 2012:

"*Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, we will have multiplied the intelligence, the human biological machine intelligence of our civilization a billion-fold.*"

Ginni Rometty, 2015:

"Some people call this artificial intelligence, but the reality is this technology will enhance us. So instead of artificial intelligence, I think we'll augment our intelligence."

Barack Obama, 2016:

"*We've been seeing specialized AI in every aspect of our lives, from medicine and transportation to how electricity is distributed, and it promises to create a vastly more productive and efficient economy … But it also has some downsides that we're gonna have to figure out in terms of not eliminating jobs. It could increase inequality. It could suppress wages*."

Vladimir Putin, 2017:

"*Artificial intelligence is the future, not only for Russian but for all of humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world*."

Andrew Ng, 2017:

"*Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI (Artificial Intelligence) will transform in the next several years*."

Geoffrey Hinton, 2018:

8 https://www.nytimes.com/2017/07/20/business/china-artificial-intelligence.html

"*I have always been convinced that the only way to get artificial intelligence to work is to do the computation in a way similar to the human brain. That is the goal I have been pursuing. We are making progress, though we still have lots to learn about how the brain actually works*. "

Many bold claims are about the emergence of AI superintelligence (ASI) with predictions of Elon Musk, Stephen Hawking, Yuval Noah Harari, Ray Kurzweil, and others. The leading scientists, including Yann LeCun, Fei-Fei Li, and Andrew Ng consider that AI will trigger a fourth industrial revolution becoming part of every aspect of human lives and extension on our capabilities.

The CEOs of the leading IT companies, Larry Page of Alphabet, the parent company of Google, and Ginni Rometty of IBM, and many others are mainly focused on expanding their business and developing new products and services based on AI. Companies avoid debate on AGI and ASI or long-term AI development.

Political leaders, including Barack Obama and Vladimir Putin, consider AI as a revolutionary technology that will bring vast opportunities and influence world politics.

Scientists and thinkers are debating, what else is needed to achieve AGI or ASI? Will DL be superseded by a different AI paradigm? Gary Markus as one of the DL sceptics is in favour of hybrid models that would incorporate not just supervised forms of DL, but other techniques as well, such as symbol-manipulation, and unsupervised learning[9]. Judea Pearl, a pioneering figure in AI, expects causal reasoning to provide machines with human-level intelligence (Pearl & Mackenzie 2018). Instead of the mere ability to correlate fever and malaria, machines need the capacity to reason that malaria causes fever. Once the causal framework is in place, it becomes possible for machines to ask counterfactual questions — to inquire how the causal relationships would change given intervention — which Pearl views as the cornerstone of scientific thought. Kahneman makes ground-breaking discoveries and explains the two cognitive systems named System 1 (Fast) and System 2 (Slow). System 1 is fast, intuitive, and emotional; System 2 is slower, more deliberative, and more logical. Kahneman exposes the extraordinary capabilities—and the faults and biases—of fast thinking and reveals the pervasive influence of intuitive impressions on our thoughts and behaviour (Kahneman 2011). Stuart J. Russell asserts the risk to humanity from advanced AI despite the uncertainty surrounding future progress in AI (Russell 2019).

Rodney Brooks demystifies the AI hype[10] and makes relevant claims why predicting the AI future is difficult, especially with over and underestimating and cites the famous Amara law:

> "*We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."*

Amara law presents a pattern found in many emerging technologies. A big promise up front, disappointment, and then slowly growing confidence beyond where the original expectations were aimed. Many technologies are overestimated in the short run, while gains are accumulated in the long run.

Amara law resembles the hype cycles[11] introduced by Gartner. The Hype Cycle is a graphical depiction of a common pattern that arises with each new technology or other innovation. Although many of Hype Cycles focus on specific technologies or innovations, the same pattern of hype and disillusionment applies to higher-level concepts such as IT methodologies and management disciplines. In many cases, the Hype Cycles also position higher-level trends and ideas, such as strategies, standards, management concepts, competencies and capabilities.

Older scientists who remember the last AI winter(s) first-hand, understand the risk of history repeating due to overpromises. Many of them are actively working in demystifying AI without provoking additional hype and framing the progress in reasonable terms.

4.2 Investments

Private companies are primary investors in the current AI period. Major internet companies invested substantial funds and hired the leading scientists, e.g. Geoffrey Hinton in Google, Yann LeCun in Facebook. Microsoft invested $1 billion in OpenAI[12]. Google acquired DeepMind for $500 million in 2014. DeepMind received investments of $570 million in 2018, $341 million in 2017, and $154 million in 2016[13]. Other companies, e.g.,

9 https://medium.com/@GaryMarcus/in-defense-of-skepticism-about-deep-learning-6e8bfd5ae0f1

10 http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/

11 https://www.gartner.com/en/documents/3887767/understanding-gartner-s-hype-cycles

12 https://openai.com/blog/microsoft/

13 https://www.forbes.com/sites/samshead/2019/08/07/deepmind-losses-soared-to-570-million-in-2018/ - 271ba0363504

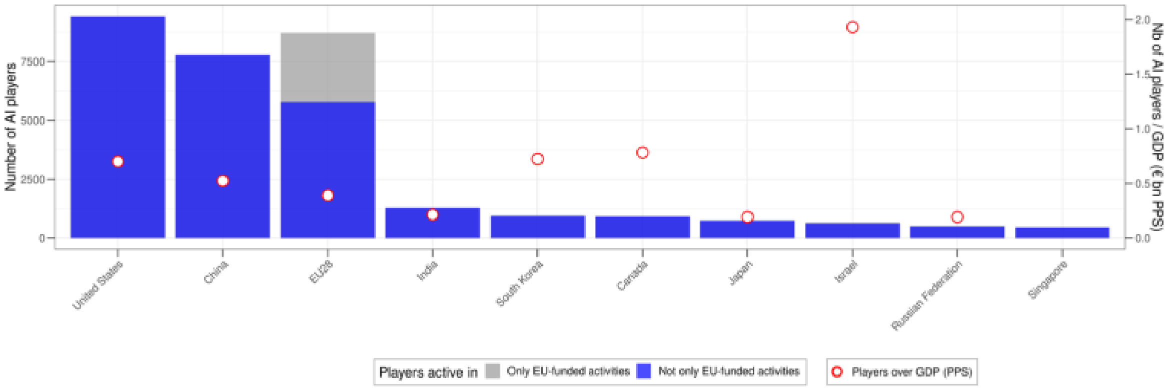
Apple, Amazon, Uber, Tesla, etc. have followed similar investments and hiring strategies. From 2013 to 2017, the amount of annual funding by Venture Capital firms increased by 350%, while all VC funding increased by 100% across all funding stages (Perrault et al. 2019).

China's ICT ecosystem is also strong, with prime leaders being Tencent, Baidu, and Alibaba, followed by Huawei. Like their USA counterparts, most Chinese companies have announced AI strategic plans with substantial investment plans. The Chinese government State Council plans for AI outline an ambitious three-stage process towards achieving China's goals on AI (Ding 2018):

- By 2020, China plans to make the AI industry "inline" with the most advanced countries. China's core AI gross industry output to exceed RMB 150 billion (EUR 20 billion) and AI output to exceed RMB 1 trillion (EUR 135 billion).
- By 2025, China aims to reach a "world-leading" level in some AI fields. The AI industry gross output to exceed RMB 400 billion (EUR 54 billion) and AI output to exceed RMB 5 trillion (EUR 680 billion).
- By 2030, China seeks to become the world's "primary" AI innovation centre. The AI industry gross output to exceed RMB 1 trillion (EUR 135 billion) and AI production to exceed RMB 10 trillion (EUR 1.35 trillion).

The European Commission (EC), in coordination with the EU Member States, has dedicated substantial AI investment funds. Under the next multiannual financial framework, the EC has proposed to dedicate at least EUR 1 billion per year from Horizon Europe and the Digital Europe Programme to AI. EU AI investment levels are low and fragmented, relative to other parts of the world, such as the US and China. The goal is to increase investments progressively to EUR 20 billion per year in the next decade. The Digital Europe program, with an overall budget of €9.2 billion, will support the digital transformation of Europe's society and economy (COM (2018) 795). Some of the EU member states have allocated significant AI national budgets, e.g., France, with €1.5 billion[14]. The European Commission plans to spend approximately €5.4 billion between 2018 and 2022, as shown in Table 3. The Member States will match this amount with over €2.6 billion. The key role of AI Watch is to monitor EU member states' AI strategies and investments and assess their impacts in terms of adoption and use of AI.

Figure 7: AI players



Note: To allow cross-country comparability, national currencies are converted into euro Purchasing Power Standard (PPS), a unit based on current euros, to account for the effect of differences in price levels across countries and of movements in exchange rates.

Source: (Samoili, Righi, et al. 2020)

One of the key early reports of AI Watch has been to analyse the landscape of Ai players worldwide. In this study, 34,000 AI players are identified worldwide in the period 2009-2018, with the US leading the landscape in number of players, mainly due to their strong industrial ecosystem (96% of US players are firms) (Samoili, Righi et al., 2020). The EU is among the geographical areas with the highest number of players active in AI, together with the United States and China (Fig. 7). The US is specialised in the areas of AI Services and in Robotics and Automation and has a leading or strong presence in all other AI thematic areas. China is the first geographical area regarding number of players filing patent applications. China's specialisation reflects its leadership in Computer vision, as well as in Machine learning and Connected and Automated vehicles. The EU hosts a very strong research network and is second by number of research institutions active in AI worldwide

14 https://www.politico.eu/article/macron-aims-to-drag-france-into-the-age-of-artificial-intelligence/

and first by number of research institutions with AI publications. Robotics and Automation, and AI Services are the EU's main areas of specialisation where it has a strong revealed comparative advantage. Other countries important in the AI landscape are India, South Korea, Canada, Japan, Israel, Russia and Singapore.

The MIT Technology Review Insights performed a survey among senior executives covering 2,357 cases and a wide range of industries worldwide (MIT insights 2018). The main conclusions are:

- 81% believe AI will have a positive impact on their industry in the future.
- 64% are likely to consider investing in AI solutions in the future.
- 78% have been asked to explore new technologies (like AI) for their business.
- 83% agree that AI will be a game-changer and significantly enhance processes across industries.

The vast majority (over 85%) of respondents consider data as the primary resource for optimal business decisions, delivering better results to customers, and growing their business.

European Commission made a survey of 9640 enterprises in January – March 2020 and measured five KPIs: AI awareness, AI adoption, AI sourcing, external and internal obstacles to AI adoption (Kazakova et al. 2020). The main conclusions are:

- Awareness of AI is high across the EU (78%). Four in ten (42%) enterprises have adopted at least one AI technology, 25% have adopted at least two. While 18% have plans to adopt AI in the next two years, 40% have neither adopted AI nor plan to do so. Adoption at the level of each technology is still relatively low: from 3% for sentiment analysis to 13% for anomaly detection and process/equipment optimisation. The most common sourcing strategy is external, as 59% of EU enterprises that use AI purchase software or ready-to-use systems.
- Three key internal barriers to AI adoption are difficulties in hiring new staff with the right skills (57%), the cost of adoption (52%) and the cost of adapting operational processes (49%). Reducing uncertainty can be beneficial, as enterprises find liability for potential damages (33%), data standardisation (33%) and regulatory obstacles (29%) to be major external challenges to AI adoption.

4.3 Issues

Deep learning has multiple issues that need to be addressed including adversarial attacks, generation of deepfake content, fairness, accountability, transparency and other ethical considerations. Here we will just briefly mention several issues, their impact and possible solutions.

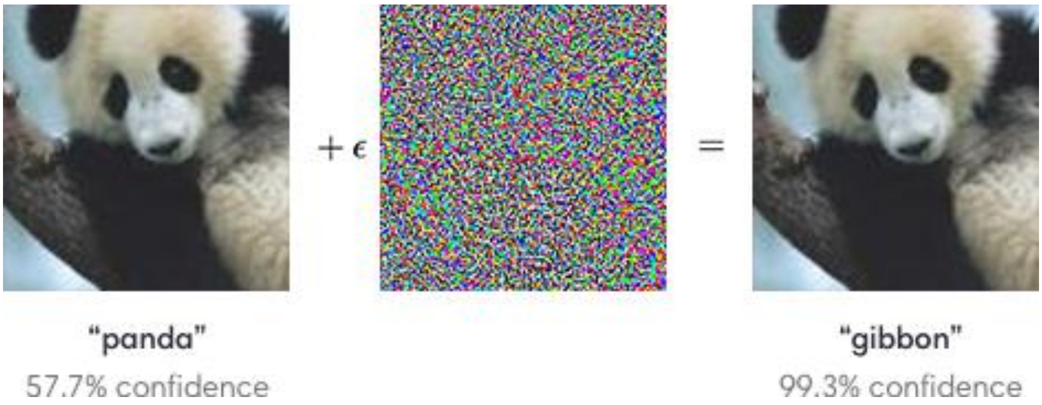
4.3.1 Adversarial attacks

Adversarial machine learning is a technique that tries to fool models by supplying deceptive input. This is probably the most common cause of a malfunction in an ML model. Most ML techniques were designed to work on specific problem sets in which the training and test data are generated from the same statistical distribution. In the real world, adversaries may supply data that violates that statistical assumption. Attackers can arrange the data to exploit specific vulnerabilities and produce massive fail in ML model.

DL is vulnerable to adversarial examples that are malicious inputs modified to yield erroneous machine model outputs. Figure 8 shows a famous adversarial example where adding a small perturbation to a panda image makes an image recognition system fail. Adversarial attacks can be very dangerous e.g., attackers can target autonomous vehicles with stickers or paint that will create an adversarial stop sign (Papernot et al. 2016).

Adversarial attacks represent an essential problem for AI safety despite significant research progress.

Figure 8: Adversarial attacks



Source: https://openai.com/blog/adversarial-example-research/

4.3.2 Deepfake

AI is reducing the cost of generating fake content and waging disinformation campaigns. The public will need to become more sceptical about online content, including text, speech, images, and videos. AI can:

- impersonate others online,
- generate fake images and news,
- automate production of fake content often posted on social media,
- automate creation of spam/phishing content.

Deepfake videos and images can depict a person saying things or performing actions that never occurred. Malicious actors with fake accounts have already begun to target the online community. For instance, it is reported that Mona Lisa can speak with only one image (Zakharov et al. 2019), or that criminals have impersonated CEOs' voice and stole millions (Kaveh et al. 2019).

4.3.3 AI ethics

Over the last years, researchers and practitioners have expressed their concerns on the ethical implications of AI systems in real-world scenarios. The European Commission appointed the High-Level Expert Group (HLEG) on AI in 2018 and one of their tasks has been to define ethical guidelines for trustworthy AI. This section is taken from the "Ethics Guidelines for Trustworthy AI" (HLEG 2019). For an AI system to be trustworthy, it should ensure the following three components through the system's entire life cycle:

1. Lawful, complying with all applicable laws and regulations,
2. Ethical, ensuring adherence to ethical principles and values,
3. Robust, both from a technical and social perspective, since even with good intentions, an AI system can cause unintentional harm.

The four ethical principles which are rooted in fundamental rights that must be respected to ensure that AI systems are developed, deployed, and used in a trustworthy manner are:

- Respect for human autonomy,
- Prevention of harm,
- Fairness,
- Explainability.

Many of these are already reflected in existing legal requirements for which mandatory compliance is required and hence fall within the scope of lawful AI, which is the first component of Trustworthy AI. As set out above, while many legal obligations reflect ethical principles, adherence to ethical principles goes beyond formal compliance with the existing laws. They are specified as ethical imperatives, such that AI practitioners should always strive to adhere to.

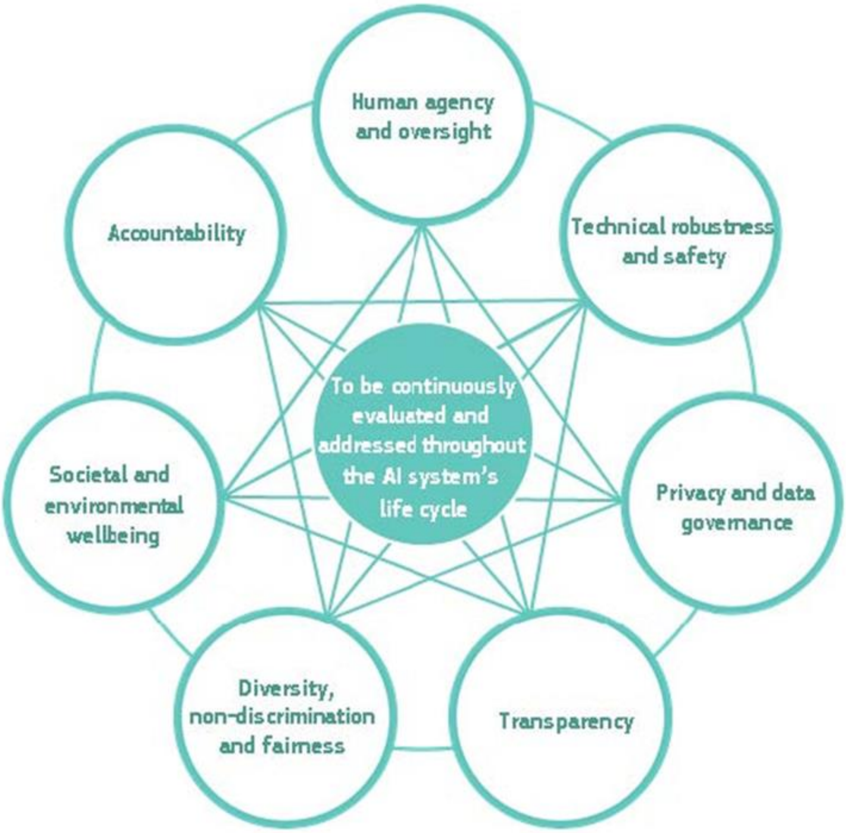
Different groups of stakeholders have different roles to play in ensuring that the AI requirements are met: *(a)* Developers should implement and apply the requirements to the design and development processes; *(b)* Deployers should ensure that the systems they use, and the products and services they offer, meet the requirements; and *(c)* End-users and the broader society should be informed about these requirements and be able to request that they are upheld.

Figure 9 presents a list of requirements that include systemic, individual, and societal aspects explained below.

1. Human agency and oversight, including fundamental rights, human agency, and human oversight.
2. Technical robustness and safety, including security and resilience to attacks, fall back plan and general safety, accuracy, reliability, and reproducibility.
3. Privacy and data governance, including respect for privacy, quality, and integrity of data, and access to data.
4. Transparency, including traceability, explainability, and communication.
5. Diversity, non-discrimination, and fairness, including the avoidance of unfair bias, accessibility, and universal design, and stakeholder participation.
6. Societal and environmental wellbeing, including sustainability and environmental friendliness, social impact, society and democracy.

Accountability, including audibility, minimization, and reporting of negative implications, trade-offs, and redress.

Figure 9: Interrelationship of the seven requirements: all are of equal importance and related to each other



Source: (HLEG 2019)

While all requirements are of equal importance, context, and potential tensions between them will need to be considered when applying them across different domains and industries. Implementation of these requirements should occur throughout an AI system's entire life cycle and depends on the specific application. While most requirements apply to all AI systems, special attention is given to those directly or indirectly affecting individuals. Therefore, for some applications (for instance, in industrial settings), they may be of less relevance.

The discussion of AI ethical principles is ongoing with the EU having a lead in the world. Following the EU, China has released its own AI principles named Beijing AI principles[15] on AI research, development, use, governance, and long-term planning. OECD produced council recommendations on the AI[16].

Google AI principles[17] focus on building socially beneficial AI, avoid creating or reinforcing unfair bias, safe, accountable to people, privacy by design, and uphold high standards of scientific excellence. Google commits not to pursue technology that will cause or likely cause harm, including AI weapons, surveillance that violates internationally accepted norms, and those that are contrary to widely accepted international laws and human rights.

Asilomar AI principles[18] in 2017 were proclaimed by leading researchers from academia and industry, and thought leaders in Economics, Law, Ethics, and Philosophy and were then embraced by an impressive list of 1273 AI/Robotics researchers and 2541 others. The Asilomar AI principles:

- Provide a broad frame about the research goals, funding, and connection with policy.
- Ethics and values consider safety, transparency, judicial transparency, responsibility, human values, personal privacy, shared benefits, human control, AI arms race, etc.
- Discuss long-term issues and risk of possible superintelligence, mitigating the threat posed by AI systems, etc.

The OECD AI Principles[19] identifies five complementary values-based principles for the responsible stewardship of trustworthy AI:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
- Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

Many AI ethical committees were formed, including Stanford institute for Human-centred AI (HAI), The Alan Turing Institute, Partnership of AI, AI Now, IEEE, and others. Recently Oxford University received $188 million donations to fund research into ethics of AI[20].

Ethical concerns have risen some related research, such as the one driven by the FAT-ML[21] (Fairness, Accountability, and Transparency in Machine Learning) community and research advances are making it possible to develop evaluation frameworks for fairness, transparency, and accountability.

15 http://www.baai.ac.cn/blog/beijing-ai-principles

16 https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

17 https://ai.google/principles/

18 https://futureoflife.org/ai-principles/

19 http://www.oecd.org/going-digital/ai/principles/

20 https://edition.cnn.com/2019/06/19/tech/stephen-schwarzman-oxford-ai-donation/index.html

21 https://www.fatml.org/

5 Conclusions

Our study reflects on the similarities and differences between AI periods. The first two AI periods share a similar pattern, as they start with a scientific breakthrough, a research paradigm shift, followed by bold predictions, vast media attention, massive investments, disappointments, unfulfilled promises, and the start of an AI winter.

The similarities and differences between the three main AI periods are illustrated Table 3. Government investments were dominant in the first two periods, while industry is nowadays investing greatly in AI developments. Academia was leading basic research in the first two periods, while today some technological companies are also leading AI basic research. These companies have been attracting leading AI scientists with cutting-edge computing resources, large-scale datasets and good working conditions, often not available in academia. Governments which have recognized the potential are investing significant resources in AI. European Union (EU) and its member states have announced an ambitious AI strategy and allocated substantial funds.

Table 2: AI periods

	Period I (1950s-1970s)	**Period II (1970s-1990s)**	**Period III (1990s-2020s)**
AI main investor	Government	Government	Industry
AI basic research	Academia	Academia	Industry/Academia
AI bold predictions	yes	yes	yes
AI Paradigm shift	Foundation of AI	Symbolic	Machine Learning and Deep Learning
Computer chip transistors[22]	68.000	27.400.000	32.000.000.000
Regulation	no	no	yes
Artificial General Intelligence	no	no	no

Source: JRC

The first paradigm shift laid the foundation for most well-known AI methods and algorithms. The second paradigm shift was related to symbolic algorithms and expert systems, also known as knowledge-based systems. The increased availability of digital data and computing power shifted the last paradigm to machine learning and deep learning. We might distinguish the second AI period from the third one with the former being model-driven and the latter being data-driven.

Computers can perform trillions of calculations per second and execute AI tasks beyond the human abilities. The tasks where AI exceeds human performance include object recognition in billions of images in a few seconds, instantaneous translation of 100+ languages, identification of spam from billions of emails, playing chess, etc. Despite these impressive performances, even the most advanced AI lacks the capabilities of a four-year-old child to move and talk, understand the world, causality, and discuss abstract concepts. Decades of AI advancements are still not a substitute for a billion years of evolution.

Clearly, AI also has negative aspects and its use can create new risks. Several AI deficiencies, including data quality and representativeness, can lead to biased, non-transparent, and low-performance AI. Malicious actors can use AI to make fake news, deep fakes, perform highly automated cyber-attacks, etc. AI risks need to be addressed. Many AI ethical committees are working on AI fairness, accountability, transparency, explainability and related issues, paving the ground for government regulation.

Many bold statements often accompany AI. Some of the bold statements will possibly fail to deliver on their promises causing disappointments and a sense of disillusion to stakeholders, venture capital companies, employees, and the general public. It is therefore important to have a balanced view of both opportunities and limitations of this emerging technology. At the present time, what we can observe is a significant increase in AI publications, conferences, patents, investment, education, national strategies, etc. Considering the current trends, it is safe to foresee in the short-term future that AI will continue to grow and spread across domains although other global emergent topics e.g. COVID-19 may impact AI expansion and change public priorities.

22 Number of transistors of Motorola 68000 (1979), Pentium II Mobile Dixon (1999), AMD Epyc Rome(2019)

Complementary to AI ethical issues, data sovereignty and technological sovereignty are becoming important. The European Commission made a strategy for policy measures and investments to enable the data economy for the coming five years[23]. This data strategy is presented at the same time as the Commission's Communication on "Shaping Europe's digital future[24]" and a White Paper on artificial intelligence[25] that indicates how the Commission will support and promote the development and uptake of artificial intelligence across the EU.

The future of AI is uncertain, with chances of another AI winter or of an even greater AI summer. Given this uncertainty, it is particularly important to have in place an unbiased initiative such as AI Watch to monitor the evolution and assess its impacts over the coming years which one way or the other will see very significant changes in our digitally transformed society.

23 https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

24 https://ec.europa.eu/info/sites/info/files/communication-shaping-europes-digital-future-feb2020_en_4.pdf

25 https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

References

ALPAC, L. (1966). Machines: Computers in Translation and Linguistics. Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, 1416.

Bertram R., *Robot Research at Stanford Research Institute*. STANFORD RESEARCH INST CA, 1972.

Bostrom, N. (2016). Superintelligence: Paths, Dangers, Strategies, Reprint ed.

Brown, N., & Sandholm, T. (2018). Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. Science, 359(6374), 418-424.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Craglia M., (Ed.), Annoni A., Benczur P., Bertoldi P., Delipetrev B., De Prato G., Feijoo C., Fernandez Macias E., Gomez E., Iglesias M., Junklewitz H, López Cobo M., Martens B., Nascimento S., Nativi S., Polvora A., Sanche. I., Tolan S., Tuomi I., Vesnic Alujevic L., *Artificial Intelligence - A European Perspective.,* EUR 29425 EN, Publications Office, Luxembourg, 2018, ISBN 978-92-79-97218-5, dos:10.2760/936974, JRC113826

Dean, J., Patterson, D., & Young, C. (2018). A new golden age in computer architecture: Empowering the machine-learning revolution. IEEE Micro, 38(2), 21-29.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE.

Desruelle P. (Ed.), Baldini G., Barboni M., Bono F., Delipetrev B., Duch Brown N., Fernandez Macias E., Gkoumas K., Joossens E., Kalpaka A., Nepelski D., Nunes de Lima M. V., Pagano A., Prettico G., Sanchez I., Sobolewski M., Triaille J.-P., Tsakalidis A., Urzi Brancati M. C., (2019) Digital Transformation in Transport, Construction, Energy, Government and Public Administration, EUR 29782 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-08613-0, doi:10.2760/689200, JRC116179

Ding J., (2018). "Deciphering China's AI Dream." *Future of Humanity Institute Technical Report*. 2018.

Domingos P., (2015). "The master algorithm: How the quest for the ultimate learning machine will remake our world" Basic Books.

EC. (2018). "*Coordinated Plan on Artificial Intelligence.*" Text. Digital Single Market - European Commission. December 7, 2018. https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence. (COM(2018) 795 final)

EC. (2020). "COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS: A European strategy for data" European Commission, Brussels, 19.2.2020 COM(2020) 66 final

EC. (2020). "WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust" European Commission, Brussels, 19.2.2020 COM(2020) 65 final

Ferrucci D., (2012). "*Introduction to "this is watson"."* IBM Journal of Research and Development 56, no. 3.4: 1-1.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artificial Intelligence Review, 48(3), 397-447.

HLEG, A. I. (2019). Ethics guidelines for trustworthy AI. B-1049 Brussels.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8), 2554-2558.

Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

Kazakova, S., Dunne, A., & Bijwaard, D., (2020) European enterprise survey on the use of technologies based on artificial intelligence, Publications Office of the European Union, Luxembourg ISBN 978-92-76-20108-3, doi:10.2759/759368.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

LeCun, Y. (2019). 1.1 Deep Learning Hardware: Past, Present, and Future. In 2019 IEEE International Solid-State Circuits Conference-(ISSCC) (pp. 12-19). IEEE.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8595-8598). IEEE.

Lighthill, J., (1973). “Artificial Intelligence: A Paper Symposium.” *Science Research Council, London*.

Michie, D. (1963). Experiments on the mechanization of game-learning Part I. Characterization of the model and its parameters. The Computer Journal, 6(3), 232-236.

Minsky, M., & Papert, S. A. (1969). Perceptrons: An introduction to computational geometry. MIT press.

MIT insights. 2018. “2018 Global Report: The Artificial Intelligence Imperative” 15.

Mitchell, T. M. (1997). Machine learning.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

Moravec, H. P. (1983). The Stanford cart and the CMU rover. Proceedings of the IEEE, 71(7), 872-884.

Newell, A., Shaw, J. C., & Simon, H. A. (1962). The processes of creative thinking. In Contemporary Approaches to Creative Thinking, 1958, University of Colorado, CO, US; Atherton Press.

Pachocki J., Brockman G., Raiman J., Zhang S., Pondé H., Tang J., Wolski. F., "OpenAI Five, 2018." URL https://blog.openai.com/openai-five.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint arXiv:1602.02697, 1(2), 3.

Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic Books.

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., ... & Niebles, J. C. (2019). The AI Index 2019 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms (No. VG-1196-G-8). Cornell Aeronautical Lab Inc Buffalo NY.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Russell S., and Norvig P., (2016). *Artificial Intelligence: A Modern Approach*. Pearson Education Limited.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.

Samuel, A. L. (1960). Programming computers to play games. In Advances in Computers (Vol. 1, pp. 165-192). Elsevier.

Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F. and Delipetrev, B., AI WATCH. Defining Artificial Intelligence, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7 (online), and doi: 10.2760/382730 (online), JRC118163.

Samoili S., Righi R., Cardona M., López Cobo M., Vázquez-Prada Baillet M., and De Prato G., TES analysis of AI Worldwide Ecosystem in 2009-2018, EUR 30109 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-16661-0, doi:10.2760/85212, JRC120106.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., ... & Roomi, B. (2017). English conversational telephone speech recognition by humans and machines. arXiv preprint arXiv:1703.02136.

Schmidhuber, J. (1993). Habilitation thesis: System modeling and optimization. Page 150 ff demonstrates credit assignment across the equivalent of 1,200 layers in an unfolded RNN.

Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Computers and biomedical research, 8(4), 303-320.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of go without human knowledge. Nature, 550(7676), 354-359.

Nativi, S., Kotsev A., Petra, S.,, Pogorzelska, K., Vakalis, I., Dalla Benetta, A., Perego., A. “IoT 2.0 and the internet of transformation (web of things and digital twins)”, JRC press, in print.

Tesauro, G. (2002). Programming backgammon using self-teaching neural nets. Artificial Intelligence, 134(1-2), 181-199.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., ... & Lau, K. (2006). Stanley: The robot that won the DARPA Grand Challenge. Journal of field Robotics, 23(9), 661-692.

Turing, A. M. (1950). Computing machinery and intelligence. The Essential Turing: The Ideas that Gave Birth to the Computer Age. Ed. B. Jack Copeland. Oxford: Oxford UP, 433-64.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Vinyals O., Babuschkin I., Chung J., Mathieu M., Jaderberg M., Czarnecki W., Dudzik A., (2019). AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language (No. MAC-TR-84). MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC.

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In Proceedings of the IEEE International Conference on Computer Vision (pp. 9459-9468).

List of abbreviations and definitions

AI	Artificial Intelligence
AGI	Artificial General Intelligence
BLEU	Bilingual Evaluation Understudy Score
CNN	Convolutional Neural Networks
CMU	Carnegie Mellon University
GAN	Generative Adversarial Network
GPU	Graphic Processing Unit
DARPA	Defence Advanced Research Projects Agency
DL	Deep Learning
DNN	Deep Neural Networks
IT	Information Technology
HAI	Human-centered AI
HLEG	High Level Expert Group
LSTM	Long short-term memory
GPS	General Problem Solver
GPS	Global Positioning System
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MOOC	Massive Open Online Courses
NN	Neural Networks
SRI	Stanford research institute
SVM	Support Vector Machines
RL	Reinforcement Learning
RNN	Recurrent Neural Networks

List of figures

List of tables

KJ-NA-30221-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office
of the European Union

doi:10.2760/801580
ISBN 978-92-76-18940-4