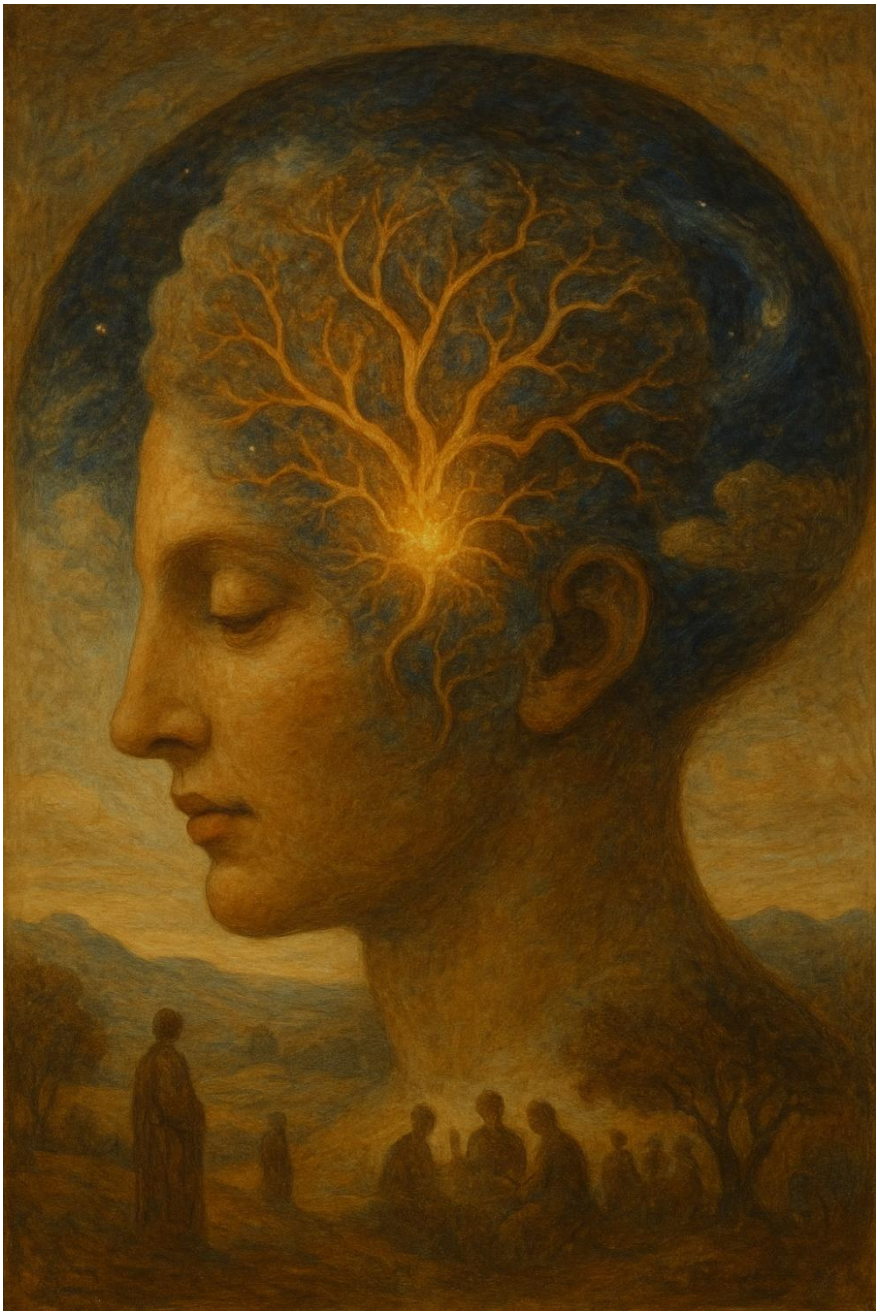
Artificial Intelligence



Part I

Technical and Economic Report, 2026

Table of Contents

1 Introduction

Artificial intelligence is rapidly transforming the economic, technological and social landscape, bringing about epoch-making changes across multiple areas. While, on the one hand, this evolution fosters innovation and production efficiency, on the other hand it raises critical issues relating to the protection of fundamental rights, the evolution of entire economic sectors, and the transformation of working, social and cultural contexts. Aware of this challenge, the European institutions, starting with the Commission, have since the previous mandate adopted initiatives aimed at addressing the phenomenon at regulatory level. In this regard, particular importance attaches to the adoption, in 2024, by the European Parliament of the Regulation on Artificial Intelligence — the so-called AI Act[1]. The Regulation, which lays down obligations concerning the use of AI systems and models on the basis of the risks associated with their use, defines different categories of risk, calibrating responsibilities and prohibitions for the actors operating within this ecosystem. Moreover, in laying down certain transparency obligations, the AI Act stresses that these are also essential for the effective implementation of the Digital Services Act (DSA — Regulation (EU) 2022/2065), emphasising the impact of the dissemination of artificially generated or manipulated content "on democratic processes, civic discourse and electoral processes, including through disinformation".[2]

The analysis offered in this report also aims to lay the groundwork for a future examination of the intersections between the DSA and the AI Act, precisely in light of the powers conferred on AGCOM as Digital Services Coordinator for Italy[3]. Accordingly, the report seeks to provide a general reconstruction of the evolution of AI in light of the role assigned to AGCOM under the DSA, without prejudice to the competences assigned by the legislator to the national competent authorities for artificial intelligence pursuant to the AI Act.

[1] See paragraph 179 of Regulation (EU) 2024/1689: the AI Act — which entered into force twenty days after its publication in the Official Journal of the EU — will start to produce its effects 24 months after 1 August 2024, with the exception of the prohibitions relating to prohibited practices (applicable from six months after the Regulation's entry into force), codes of practice (applicable from nine months after entry into force), rules on general-purpose AI systems, including governance (applicable from twelve months after entry into force), and obligations for high-risk systems (applicable from thirty-six months after entry into force).

[2] See paragraph 120 of Regulation (EU) 2024/1689.

[3] Competence assigned pursuant to Decree-Law No. 123 of 15 September 2023, converted into law with amendments by Law No. 159 of 13 November 2023 (Official Gazette No. 266 of 14 November 2023).

In view of its institutional responsibilities and in light of the ongoing technological evolution, AGCOM has therefore launched a series of activities aimed at analysing artificial intelligence, its effects on the sectors and matters falling within the Authority's institutional remit, and at monitoring developments in the European and national regulatory framework. To this end, the Authority first established, by Resolution No. 11/24/CONS of 24 January 2024, a Committee on Artificial Intelligence (hereinafter also the Committee), in order to ensure qualified and specialised support regarding the implications of AI systems for the Authority's areas of competence and the role it may assume in this field, with advisory functions.[4] Subsequently — as part of the reorganisation of the Offices — it was decided to provide the Authority with an ad hoc structure tasked with carrying out study and analysis activities in the fields of big data and AI, as well as coordinating the Committee itself (Resolution No. 382/24/CONS of 30 September 2024).

In view of the need to ensure constant oversight of the relevant technological context, this working document launches monitoring activities concerning the evolution of technologies, the regulatory framework and the debate on artificial intelligence, also with a view to fostering in-depth reflection on the related implications.

As explained in the Foreword, this document should be read together with the Report on Artificial Intelligence prepared by the Committee, which completes and further develops the analysis of various legal and regulatory issues.

[4] The Committee has a two-year term, renewable where appropriate, and is composed as follows: Prof. Andrea Renda (European University Institute), coordinator; Prof. Giovanni Boccia Artieri (University of Urbino), member; Prof. Giuseppe Cassano (European School of Economics), member; Prof. Mauro Giusto (University of Milan), member; Prof. Andrea Simoncini (University of Florence), member; Prof. Giovanna de Minico (University of Naples Federico II), member; Andrea Imperiali, member.

2 Evolution of AI

The emergence of AI as a scientific discipline can be traced back to a specific event*: the Dartmouth Summer Research Project on* Artificial Intelligence of 1956. The organisers of that conference started from the assumption that every aspect of **learning**, or any other feature of intelligence, could in principle be described so precisely that it could be simulated by a machine.[5] Those same scientists attempted to find ways to make machines use language, form concepts, solve problems — which at the time were reserved for human beings — and ultimately improve the machines themselves.[6]

That experience first led one of the participants in *the summer* school — John McCarthy — to coin the term "Artificial Intelligence" (namely: "The *science and engineering of making intelligent machines"),*[7]and subsequently to the establishment of the AI Lab at MIT. The term "Artificial Intelligence" was coined after the definition of the criterion aimed at assessing whether computers were capable of mimicking human intelligence, namely what would become known as the "**Turing test**"[8] (see also §§ 2.1 and 3).

Box 1 – Turing Test

Proposed in 1950 by Alan Turing, the test is a criterion for determining whether a machine is capable of exhibiting intelligent behaviour. The test is based on a Cartesian idea of what a human being is, and indeed its reliability has recently been questioned. It consists of a written conversation between a human being and a machine, mediated by another human being (the interrogator), whose task is to identify which of the two interlocutors is human. If the machine succeeds in deceiving the interrogator to the point of being indistinguishable from a human being, it is considered to have passed

[5]The basic idea of the event organised by John McCarthy, Marvin Minsky, Claude E. Shannon and Nathaniel Rochester (an IBM computer *scientist) was that "every aspect of learning or any other feature of intelligence can, in principle, be so precisely described that a machine can be made to simulate it". See: McCarthy,* J., Minsky, M.L., Rochester, N., Shannon, C.E., A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955: "every *aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate* it".

[6]McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., cit. The following passage from the text is reproduced: *"The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be* made *to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer".*

[7]McCarthy, John. (2007). What is artificial intelligence.

[8]Turing, A.M. (1950), Computing machinery and intelligence, *Mind,* 49.

the test (in this context thresholds, or percentages, are applied beyond which the machine is considered indistinguishable; in the classical interpretation, these start from 30%).

The main objective of the test is to demonstrate that a machine can simulate human intelligence at least at the linguistic level. However, it does not measure consciousness or genuine understanding, but only the ability to imitate human linguistic behaviour.

Today, advanced language models are able to pass the test in many contexts, but this does not necessarily imply that they possess general intelligence or self-awareness. [9][10]Consequently, the validity of the test as definitive proof of intelligence has been questioned. Since these machines are now able to imitate human register and fluency almost perfectly through sophisticated probabilistic calculations, the Turing test today tends to measure the capacity for "simulation" rather than a real form of understanding or consciousness. Passing the test is no longer regarded as the ultimate goal, since a system may appear human without possessing any real capacity for abstract reasoning or intentionality.

For this reason, the scientific community is adopting new and much more rigorous evaluation parameters. Among the most relevant are MMLU *(Massive Multitask Language Understanding),* which assesses knowledge across dozens of academic disciplines, and *ARC-AGI (Abstraction and* Reasoning Corpus). The latter, in particular, challenges models to solve novel logical problems requiring fluid intelligence rather than the mere repetition of patterns learned during training, seeking to draw a clear boundary between language imitation and genuine problem-solving ability (for an analysis of the contemporary implications of LLMs passing the Turing test, see §§ 3.2.2 for the technical 3.2.2framework and 5.5 for 5.5implications regarding the role of LLMs as an infrastructure for access to information).

2.1 First AI Spring (1956–1973)

The 1956 event generated great enthusiasm within the scientific community and was initially described as an *"AI* spring" (that is, a "spring" of artificial intelligence), now also known as the first wave of artificial intelligence. The term "wave" conveys the cyclical nature that has characterised the evolutionary phases of AI, where great expectations were followed by limited progress, leading — in certain phases known as "AI winters" — to a loss of interest among

[9]Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences,* 121(9), e2313925121. See also: "Study finds ChatGPT's latest bot behaves like humans, only better".

[10]Biever, C. (2023). ChatGPT broke the Turing test-the race is on for new ways to assess AI. *Nature,* 619(7971), 686-689.

scientists and investors in the sector and in the topic (see Figure 1). Indeed, the history of AI has seen two major setbacks: the first between 1974 and 1980, and [11]the second between the late 1980s and 2011.[12]

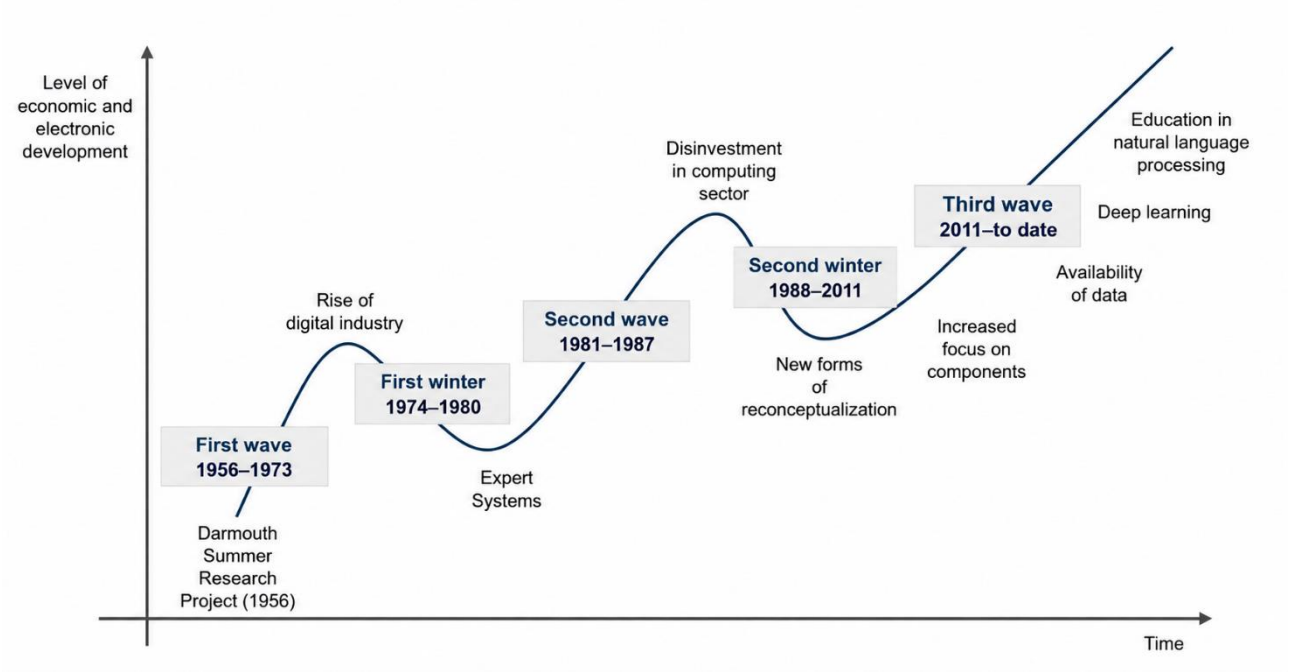


Figure 11 – AI Timeline

Returning to the origins of the discipline, interest in the Dartmouth conference triggered the development of various programs, including some capable of playing [13]checkers. Among them, the program developed by Allen Newell and Herbert Simon attracted considerable attention: they presented Logic Theorist (LT), a program capable of developing theories relating to **symbolic logic**, together with a processing language called IPL (Information Processing Language).[14] The program represents a watershed in the history of AI because it opened the

[11]Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., & Forghani, R. (2020). Brief history of artificial intelligence, *Neuroimaging Clinics of North America,* 30(4), 393-399.

[12]Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021). "A brief history of AI: how to prevent another winter (a critical review)", *PET Clinics,* 16(4), 449-469.

[13]Sheikh, H., Prins, C., & Schrijvers, E. (2023). Mission AI: The new system technology, *Springer Nature,* 410.

[14]In addition to being one of the founding fathers of the discipline, Allen Newell later contributed to the development of cognitive psychology, while Herbert Simon focused, among other things, on the concept of rationality, in particular "bounded rationality"; for these studies he was awarded the Nobel Prize in Economics in 1978.

way to the solution of conceptual and logical problems, to the point of being able to prove some of the theorems contained in Whitehead and Russell's Principia Mathematica. In 1959, Newell and Simon released the General Problem Solver (GPS), a program designed to imitate the problem-solving protocols of the human brain:[15] the software proposed solutions by translating problems into goals and sub-goals, and by identifying actions and operators.[16] Through this decoding of reality, GPS was able to solve the logical river-crossing puzzle known as the "Missionaries and Cannibals problem".[17] However, although the program worked well when applied to simple problems, the enthusiasm generated by GPS faded once researchers realised that it could not have the generalised application its name suggested.[18] In addition to the system's intrinsic limitations, it is also clear that its performance — like that of all early programs — necessarily suffered from constraints relating to the size and speed of memory and processors.[19]

In the following years, several theoretical advances were recorded, although with limited impact outside laboratories. According to John Launchbury, former Director of DARPA's Information Innovation Office, the first AI spring can be traced back to the notion of "Crafted Knowledge", a definition that includes those artificial intelligence systems based on **rules** supplied to machines by programmers.[20] Indeed, by circumscribing the complexity of an event and reducing it to the rules given by programmers, machines were capable of reasoning in relation to certain **narrow domains**, while lacking the ability to learn, to abstract, or to manage situations characterised by a certain degree of uncertainty. Within this framework came the studies of Marvin Minsky, who in 1963 proposed a simplification approach for AI use cases. Together with Seymour Papert, Minsky suggested that AI studies should focus on designing

[15]The program was in fact able to solve most of the theorems in the second chapter of the book: 38 out of 52.

[16]Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program, in *IFIP Congress* (vol. 256, p. 64).

[17]Boden, M. A. (2008). *Mind as machine: A history of cognitive science.* Oxford University Press: *"GPS could even solve the tricky missionaries-and-cannibals puzzle, which requires one to go backwards in order to go forwards (Three missionaries and three cannibals on one side of a river; a boat big enough for two people; how can everyone cross the river, without cannibals ever outnumbering missionaries?). Purpose, thinking, and mental representations: all these seemed within reach at* last".

[18]Coppin, B. (2004). *Artificial intelligence illuminated.* Jones & Bartlett Learning.

[19]Buchanan, B. G. (2005). A (very) brief history of artificial intelligence, *AI Magazine,* 26(4), 53-53.

[20]See the video produced by *DARPA's Information Innovation* Office.

programs capable of producing responses in smaller artificial environments: the so-called *"block* universe" ("micro-worlds").[21]

A few years earlier, Minsky and Papert had written a book (Perceptrons) challenging Frank Rosenblatt's work on neural networks;[22] Rosenblatt is today considered the father of neural networks for having introduced the "error-based [23][24]perceptron learning rule". With their work, the authors intended to criticise one of the approaches to the study of AI that had emerged during the first spring: adopting a **symbolic/logical** approach, they opposed **connectionism**, the paradigm underlying Rosenblatt's networks. While the symbolic/logical approach (or "symbolism") was based on the idea that intelligence could arise from symbols and logical rules, and that the learning of an AI model should rely on such rules through a mechanism of logical deduction, connectionism was based on the idea of an AI capable of simulating the functioning of the human brain by connecting artificial networks and therefore replicating the structure of neurons. Learning from data was continuous and took place through artificial neural networks. This latter approach, using the first neural networks, was clearly destined to re-emerge later in the history of AI with machine learning (see § 3.1.2) 3.1.3deep learning (see § 3.1.3). But in 1969, precisely in Perceptrons, examples of mathematical demonstrations were presented that neural networks were unable to solve; this contributed to spreading scepticism toward this approach — which would later prove successful — and to the beginning of [25]the first AI winter.

2.2 First AI Winter (1974–1980)

In the same direction as the criticisms advanced by Minsky, two reports — the first released by the US government (the ALPAC report)[26] and the second by the British government (the 1973

[21]Minsky, M., & Papert, S. A. (1972). Artificial intelligence progress report. The authors state that: "To get *experience with broader, if shallower, systems we plan to build up small models of real world situations; each should be a small but complete heuristic problem-solving system, organized so that its functions are openly represented in forms that can be understood not only by programmers but also by other programs. Then the simple-minded solutions proposed by these mini-theories may be used as plans for more sophisticated systems, and their programs can be used as starting points for learning programs that intend to improve* them".

[22]Minsky, M. L., & Papert, S. A. (1988). *Perceptrons: expanded edition.*

[23]Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project* Para. Cornell Aeronautical Laboratory.

[24]Kautz, H. (2022). "The third AI summer: AAAI Robert S. Engelmore memorial lecture", AI *Magazine,* 43(1), 105-125.

[25]Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The new system technology* (p. 410). Springer Nature. See Chapter 2, Artificial Intelligence: *Definition and Background, p. 32.*

[26]Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, and National Research Council Publication, Language *and Machines Computers in Translation and Linguistic, Publication* 1416, 1966.

Lighthill report)[27] — offered a bleak forecast regarding the prospects of technology based on artificial neural networks. This led to a drastic reduction in research support for the entire sector,[28] slowing its development until the 1980s.[29]

Among the causes of the setback in AI research, three key factors can be [30]identified:

- the fact that systems were developed with the aim of reproducing human thought: instead of adopting a bottom-up *approach* starting from an in-depth analysis of the task/problem to be assigned to the machine through the identification of a possible solution condensed into an algorithm, programmers attempted to replicate the way human beings perform a given action or solve a problem;
- **the oversimplification of problems:** the model proposed by Minsky led to an excessive simplification of reality. This reductionism appeared fruitful only when applied to programs that attempted to obtain a solution by combining a sequence of (simple) steps;
- the third factor was related to critical reflections on neural networks and the limits of their fundamental structures: as already noted, Minsky emphasised the limitations of neural networks in 1969 and this had repercussions on investment in the AI sector.

2.3 Second AI Spring (1981–1987)

The second AI spring took place in the 1980s thanks to the rise — during the previous decade — of symbolic AI, which was based on the idea that machines could be instructed by providing them with certain rules (initial data established by professionals in a specific sector), on the basis of which information would then be inferred. This is therefore referred to as the "**expert system era**" and "**rule-based AI**", that is, the attempt to provide machines with information selected by experts and relating to a specific sector, therefore in narrow areas of competence.[31] Among the most famous projects that preceded the second flowering of AI, several celebrated

[27]Sir James Lighthill was commissioned by the British Parliament to assess the state of AI research in the country. The report concluded that all experiments carried out in the field of AI would have been better handled by researchers from other disciplines, and that AI successes in "toy problems" could never be adapted to real-world applications because of combinatorial explosion. The report was published in 1973 by the Science Research Council (SRC). The first part of the report is available here.

[28]Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021), cit.

[29]In the United States, research funding was reduced.

[30]Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021), cit.

[31]Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021), cit.

"expert systems" may be recalled: DENDRAL (1968), a program capable of inferring molecular structure from mass spectrometry data; and MYCIN (1975), a rule-based program supplied by physicians to identify bacteria causing sepsis and recommend antibiotic dosage according to the patient's weight. Drawing on a set of about 600 rules, researchers found that MYCIN could make more accurate diagnoses than those provided by medical residents.[32]

Only from the 1980s, when the probabilistic component was introduced into expert systems, did these programs find wide application in the industrial and commercial sphere, becoming a crucial component of companies' research and development. In particular, the leap in the application of expert systems occurred with DEC's commercial implementation [33]of XCON. XCON significantly accelerated system configuration: whereas in the 1970s programming a computer system was a slow and error-prone process, XCON reduced the time required to generate a satisfactory system configuration to about 90 minutes. Thanks to this innovation, companies invested in software and hardware ecosystems, while software start-ups offered *expert system shells* on the market, equipped with user interfaces that enabled even non-programmers to input rules. The programming languages used in experiments related to artificial intelligence were LISP (created by McCarthy 20 years earlier) and Prolog.[34]

These computer developments meant that the second AI spring differed from the first because investment was driven by pressure from both companies and governments. However, despite the impetus from corporations and governments, especially the American and Japanese governments, the initially hoped-for results were never achieved (IBM introduced to the market a PC more powerful than those specifically designed for AI, such as LISP machines), and this led the sector to face the second AI winter, a setback that lasted until 2011.[35]

2.4 Second AI Winter (1988–2011)

As early as 1984, four years before the beginning of the second winter, during the annual meeting of the Association for the Advancement of Artificial Intelligence (AAAI), Roger Schank

[32]Shortliffe, E. H., & Buchanan, B. G. (1975). "A model of inexact reasoning in medicine", Mathematical *Biosciences, 23(3-4),* 31-379.

[33]Polit, S. (1984). "R1 and beyond: AI technology transfer at digital equipment corporation", AI *Magazine,* 5(4), 76-76.

[34]Kautz, H. (2022), cit.

[35]Ibid.

and Marvin Minsky predicted the onset of a second "AI winter" and the consequent **collapse of investment** in the field of AI, with a reduction in funding similar to that which had occurred in the mid-1970s.[36] But despite the contraction in AI investment and the consequent slowdown of the sector in terms of research and development (the limited commercial interest in AI applications resulted from the failure of expert systems and of the Japanese Fifth Generation Computer Systems hardware), programmers continued to elaborate on what had been produced in previous decades. Indeed, what the literature defines as the "second AI winter" was in fact a period marked by a decline in available funding and a fall in public interest in the subject, but this did not coincide at all with a halt in scientific research. Scientific efforts meant that the second AI winter cannot be considered a sterile period from the point of view of research. On the contrary, as outlined below, some of the discoveries made during this period laid the foundations for today's explosion of AI applications.

In this sense, the attempt to revisit the backpropagation algorithm was remarkable. Backpropagation is the primary learning mechanism used to train artificial **neural networks** (on the concept of backpropagation, also § 3.1.3). Initially conceived by Arthur Bryson and Yu-Chi Ho in 1969 as a method for optimising dynamic systems,[37] it was later taken up — in 1986 — by Hinton, Rumelhart and Williams, who demonstrated that the algorithm *could* effectively train multi-layer neural networks in order to find solutions to non-linear problems.[38] As will be seen below, backpropagation would be of fundamental importance in the third AI spring, as it was crucial for solving machine-learning problems and became a cornerstone of neural networks. Thanks to this element, and to the increase in computational power in the early 2000s, researchers were able to use neural networks to overcome the learning limitations shown by symbolic AI systems in fields such as image recognition and speech processing, making backpropagation the standard training method.

Before the return to neural networks during the third AI spring, however, in 1995 — with the introduction of support-vector machines (SVMs) — classical machine learning reached its

[36]In this regard, see the article: "Marvin Minsky and Roger Schank warned of a second AI winter in 1984".

[37]Bryson, A. E., Ho, Y.-C. (1969), Applied *optimal control, Routledge,* 2018.

[38]Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-Propagating Errors, *Nature,* 323(6088), 533-536.

highest *point ("the most powerful approach to 'black box' machine learning".* On this topic, see also § 3.1.2).[39]

As mentioned, in the history of AI, various artificial intelligence projects had sought to condense knowledge into languages (statements composing formal languages) so that machines could perform automatic processing on the basis of a set of logical inference rules (symbolic AI; see "symbolism", § 2.2). However, it already been pointed out that this approach — based on the use of symbols and formal rules — although it played a fundamental role in the history of AI, did not achieve great success,[40] and that **connectionist theory** ("connectionism", § 2.2), despite its initial difficulties in terms of computing speed and data availability, has today prevailed (particularly from 2012 onwards, with the advent of deep learning. On point, see also § 3.1.3).

Indeed, as noted by several authors (including Yoshua Bengio, who in 2018 would win the Turing Award for deep networks):

> *"A person's everyday life requires an immense amount of knowledge about the world. Much of this knowledge is subjective and intuitive, and therefore difficult to articulate in a formal way. Computers need to capture this same knowledge in order to behave in an intelligent way. One of the key challenges in artificial intelligence is how to get this informal knowledge into a computer".* [41]

It was therefore highlighted that many of the successes recorded in the field of AI up to that point had taken *place "in relatively sterile and formal* environments", where it was not necessary for the machine to have extensive knowledge of the surrounding world. Among these successes was undoubtedly IBM Deep Blue's victory over world champion Garry Kasparov in 1997. Since the rules of chess are confined to a very short and formal list, this success too falls within knowledge-based AI.[42] In this regard, some authors stress that Deep Blue's performance — precisely because of the set of rules defined by chess players and encoded within it by programmers — should not be interpreted as the victory of the intelligent computer over the

[39]Kautz, H. (2022), cit.

[40]Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[41]Ibid.

[42]Ibid.

human being, but rather as the collective triumph of a computer and countless (human) players over a single grandmaster. [43]

2.5 Third AI Spring (from 2012 to the present)

Deep Blue's victory rekindled interest in AI (IBM's share price increased tenfold), [44]and this accelerated research on neural networks, also in relation to the evolution of studies on the backpropagation algorithm (mentioned above, § 2.42.4), and therefore on the training of multi-layer neural 3.1.3[45]networks (see § 3.1.3) and on the identification of patterns by machines.

Two factors were essentially responsible for enabling the **advancement of research on neural networks** and allowing AI to acquire its own capacity to perceive the natural world:

- the availability of **enormous amounts of data (big data)**, especially unstructured data (texts, images, videos, etc.), some of which were labelled (labelled data), [46]now abundantly available thanks to the development of platforms;
- the increase in **computational power** (the increase in the computing power of GPUs — Graphics Processing Units), which made it possible, through the use of available data, to train deeper networks and larger models in increasingly shorter times. [47-48]

These two factors, which have progressively reinforced each other, have given a huge boost to the applicability of neural networks in the field of machine learning. In substance, the enormous and growing availability of data of all kinds, combined with increased computational capacity,

[43]Sheikh, H., Prins, C., & Schrijvers, E. (2023), cit. See Chapter 2, *Artificial Intelligence: Definition and* Background, p. 35.

[44]Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021), cit.

[45]Sheikh, H., Prins, C., & Schrijvers, E. (2023), cit. See Chapter 2, *Artificial Intelligence: Definition and* Background, p. 35.

[46]One of the most famous data-labelling *projects* was ImageNet, launched in 2007 by Stanford University. The project contained more than 14 million manually labelled images. Each image was associated with one or more labels identifying the objects present. The project had a fundamental impact on the progress of deep *learning*, particularly thanks to the *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC), an annual competition launched in 2010. In particular, in 2012, when Hinton and Krizhevsky's team won the competition with AlexNet, a deep convolutional neural network that far outperformed previous results, that moment is often identified as the beginning of the modern era of deep *learning*.

[47]Eeckhout, L. (2017). Is Moore's law slowing down? What's next?. IEEE Micro, *37(04),* 4-5.

[48]Today Moore's law appears to have declined, giving way to what is referred to as Huang's law (after the CEO and co-founder of Nvidia). In this regard, see Perry, T. S. (2018). Move over, Moore's law. Make way for Huang's law [Spectral *Lines]. IEEE* Spectrum, 55(5), 7-7. *Specifically: "Just how fast does GPU technology advance? In his address, Huang pointed out that Nvidia's GPUs today are 25 times as fast as they were five years ago. If they were advancing according to Moore's Law, he said, they would have increased* in speed by only a factor *of 10." See also Hao, K. (2019), The computing power needed to train AI is rising seven times* faster than before, MIT Technology Review.

has triggered a virtuous process (also a positive feedback loop) that has made possible the radical improvement of learning-network performance and therefore the definitive establishment of artificial intelligence, not only as a research tool but also as a service to the public. Consequently, rather than speaking of a third spring, we should speak of the **affirmation of AI** in all economic and social fields.

Through the use of multiple layers, the problem whereby models trained on existing data were unable to process new information effectively was overcome. The use of multiple layers in the training process has taken the name "deep learning": each layer provides a more complex representation of the input than the previous one. A practical example in image recognition: while the first layer may be able to identify corners and points, the second can distinguish parts of a face such as the tip of a nose or the iris of an eye; the third layer is able to recognise whole noses and eyes, and so on until reaching a layer that recognises the face of a single person (see 3.1.3).

In 2014, Ian Goodfellow introduced generative adversarial networks (Generative Adversarial Networks — GANs): a deep learning architecture in which two neural networks are trained to compete against each other in order to generate new and more genuine training datasets having the same distribution as the data initially supplied during training.

Although artificial intelligence was already used in many digital services (for example online translators), public attention to deep neural networks was renewed in 2016, when Google DeepMind's AlphaGo defeated the world Go champion. One year later, [49]Google researchers presented the paper "Attention Is All You Need", which defined a new deep learning architecture called the transformer (see § 3.2 for some technical details on transformers).

On the basis of these innovations, in 2018 OpenAI launched the GPT project (which has now reached 4th version), [50]whose acronym stands for Generative Pre-trained Transformer. Its mass-scale diffusion and impact on public debate, however, emerged above all from 2022 onwards, with the rapid adoption of conversation-based services built on LLMs and the emergence of

[49]Go is a board game for two players, who alternately place black and white stones on the empty intersections of a board formed by a 19 × 19 grid. The aim of the game is to control a larger area than that controlled by the opponent. Go originated in China, where it has been played for at least 2,500 years; it is very popular in East Asia and has spread to the rest of the world in recent years.

[50]ChatGPT was launched globally on 30 November 2022, making the service accessible to users in Italy from that date.

several competing actors (for the market implications connected with foundation models and LLM services, including dynamics of integration and concentration, see Chapter 4).

With the aim of producing an algorithm capable of interacting with humans in the most natural possible way, OpenAI adopted a large pre-trained language model (so-called LLMs, i.e. Large 3.2.2Language Models; see also § 3.2.2), introducing a catalytic element for natural [51]language processing (NLP). GPT — which [52]uses the decoder component of the transformer — offered a different approach to pre-training, focusing on the generation of coherent text tailored to the user's [53]context through autoregressive modelling.

This model is based on a dialogic service, defined as a "chatbot", that is, a machine ("bot" from robot) capable Box 2sustaining a conversation (see Box 2), establishing a direct and personalised relationship with the user. This anthropomorphisation of chatbots — which operate on the basis of LLM models — requires an ever-increasing amount of human-generated data. Models trained on large amounts of data are called foundation models. The data are used to refine the parameters (i.e. the weights and orientations that constitute the models' internal logic), thus enabling AI to understand the prompts formulated by users — namely the instructions [54]and questions addressed to the machine — and to provide more accurate responses on the basis of a 3.1.3probabilistic association (for an overview of these models, see § 3.1.3).

Box 2 – Chatbot

The idea of an intelligent machine capable of conversing with humans is a suggestion that has always characterised the development of AI. As early as the 1950s, with the test that bears his name, Alan Turing introduced a criterion for assessing whether a

[51]Unlike Large Language *Models (LLMs), which* have hundreds of billions of parameters, Small *Language Models (SLMs)* are designed to offer advanced reasoning capabilities with a fraction of the computational power required. This recent trend responds to the need for greater energy sustainability, lower operating costs and the possibility of running AI locally on devices *(on-device),* thereby ensuring greater data privacy (see Chapter 3).

[52]Decoder-only *LLMs are able* to translate text, but also to generate it, and therefore to create content. GPT-1 used only the decoder component (with 12 layers) of the transformer architecture.

[53]The GPT-1 model could be classified as *semi-supervised learning* because it was characterised by an unsupervised pre-training phase and a supervised fine-tuning phase.

[54]A prompt is a natural-language request sent to a language model in order to receive a response.

machine could be considered "intelligent" on the basis of its ability to sustain a conversation indistinguishable from that of a human being (see §§ 2.2 and 3.3).

The first chatbot in history was then created in 1966, when Joseph Weizenbaum, a researcher at MIT, developed ELIZA. ELIZA simulated a psychotherapist by using the words employed by the user to formulate its own questions, without however truly understanding the meaning of the terms used.

Less than ten years later, in 1972, PARRY was created: a more sophisticated chatbot capable of simulating a person with paranoid schizophrenia through a model able to simulate emotional reactions.

In the 1980s, AI began to be applied in "expert systems", programs designed to simulate human reasoning in specific fields, albeit with limited success.

Between the late 1990s and the early 2000s, with the advent of the Internet and progress in natural language processing (NLP), more advanced chatbots emerged, such as A.L.I.C.E. (Artificial Linguistic Internet Computer Entity, capable of using more complex patterns) and SmarterChild, a precursor of modern virtual assistants. SmarterChild — introduced in 2001 — was available on MSN Messenger and AOL Instant Messenger and was designed for entertainment and information, being able to provide responses on news, weather and other services.

From 2010 to the present, with the integration of AI into digital services and of those services into products that enable an easier relationship between machines and users, virtual assistants have become established (such as Apple's Siri in 2011 or Amazon Alexa in 2014). In parallel, more and more companies have begun to develop chatbots for customer service and communication automation.

Finally, with the introduction of deep *learning models,* chatbots have become increasingly sophisticated and capable of understanding and generating text in an increasingly natural way.

With 175 billion parameters, GPT-3 was an innovative model in the LLM landscape because it was able to generate human-like text, responding to users' prompts with minimal fine-tuning. GPT [55]represented a turning point thanks to its size and its ability to perform a wide variety of

[55]GPT-3 was in fact able to pool what it had learned, using such "generalised" knowledge to solve problems in different domains without the need to modify the model parameters. In other words, thanks to a very large number of parameters, GPT-3 enabled the model to handle a wide range of activities, from text generation to translation, without requiring specific training for each type of requested task. Having expanded the model's multimodal capabilities, in 2023 OpenAI released GPT-4. This evolution meant that the model could accept not only text but also images as input, thereby allowing it to understand and interpret visual information as well as natural language. In addition, GPT-4 significantly improved the model's ability for complex reasoning and contextual understanding, enabling it to produce detailed summaries and extended dialogues, with reasoning more similar to human reasoning and a reduced presence of bias. Unlike GPT-3 (175 billion parameters), no complete official data are available for GPT-4 on size and architecture; heterogeneous estimates circulate in the literature and public debate. In general, the observed leap in capabilities is attributed not only to scale (parameters and data), but also to the evolution of training and alignment procedures (pre-training, supervised fine-tuning and RLHF techniques), as well as

linguistic tasks without specific training for each task, showing the general public the potential and versatility of artificial intelligence as it is used today through its most widespread applications.

Although GPUs (Graphics Processing Units) represented the fundamental catalyst for the affirmation of deep learning, the current phase of development is characterised by a push towards even more specialised hardware. The shift has been from processors originally designed for graphics to circuits designed exclusively for AI, such as TPUs (Tensor Processing Units), optimised to accelerate tensor operations typical of neural networks. At the same time, the spread of artificial intelligence on commonly used devices has led to the integration of NPUs (Neural Processing Units) into smartphones and PCs. This evolution is aimed not only at increasing computing speed but above all at energy efficiency and latency reduction, making it possible to run complex models directly on the user's device (on-device AI) without constantly depending on cloud servers.

In this context, new AI models, starting with OpenAI's ChatGPT-4, exhibit *"behavioral and personality traits that are statistically indistinguishable from a random human from tens of thousands of human subjects from more than 50 countries".*[56] In substance, modern generative artificial intelligence services have gone beyond the requirements that Alan Turing and his colleagues set themselves after the war (see Box 2), not only making it difficult to distinguish them from human behaviour, but also displaying personality traits that are sometimes "more human" than what we consider human.

architectural and engineering choices. However, although fine-tuning required fewer resources than training a model from scratch, large transformers such as GPT-4 still required considerable computing power because of the enormous volume of data used. In May 2024, GPT-4o was released, representing the latest advance in OpenAI's language models. With the new version, OpenAI introduced improvements that now return more concise answers and more structured scientific explanations, while showing improved creative-writing capabilities; over time, however, the model has been joined by a variety of competitors.

[56]Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences,* 121(9). It should also be considered that, although these artificial intelligence systems have shown that they can pass the Turing Test, many experts argue that the test focuses excessively on the ability to simulate human conversation, overlooking other forms of intelligence. For a discussion of this topic, see: Sejnowski, T. J. (2023). Large language models and the reverse Turing test. Neural Computation, *35(3), 309-342;* Jones, C., & Bergen, B. (2024). Does GPT-4 pass the Turing test? In Proceedings *of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1,* pp. 5183-5210; Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish GPT-4 from a human in a Turing test. *arXiv preprint* arXiv:2405.08007.

2.6 Concluding remarks: AI models, drivers and actors

Over the last seventy years, artificial intelligence has developed through a series of profound transformations concerning not only the technologies used, but also the motivations that guided research and the actors involved in its construction (see Table 1).

Wave	Decades	Models	Drivers	Main actors
First	1950–1970	Rule-based systems, symbolic logic *(Crafted* Knowledge)	Scientific	Public model: scientists and researchers (universities and state funding)
Second	1980	Expert systems	*Classified knowledge*	Mixed model: cooperation between the State (e.g. DARPA) and private companies
Third	2010–present	Learning models (machine *learning and* deep neural networks/deep *learning)*	Big data/computational capacity	Private model: global digital platforms

Table 1– Evolution of AI: models, drivers and actors

From the point of view of **models**, AI has moved from symbolic approaches — based on rules and explicit representations of knowledge — to techniques increasingly oriented towards learning from data (machine learning and deep learning). Whereas the first generations of systems were built "by hand" by experts, the most recent ones are based on statistical models and deep neural networks, capable of learning autonomously through exposure to large amounts of data. This transition marks an epistemological turning point: from "programmed" intelligence to "learned" intelligence (§ 3.1).

Box 3 – Evolution of the ethical debate

The ethical debate on AI has undergone a transformation parallel to its technical evolution. If in the early phases (1950–1980) reflections were often confined to philosophy or science fiction — consider, for example, the so-called "Three Laws of Robotics" by the writer Isaac Asimov — the advent of

the Third Spring has made these challenges extremely concrete and urgent. Today attention focuses on four fundamental pillars (see Chapter 5):

- **Bias and discrimination:** Since models learn from existing data, they risk reflecting and amplifying the human biases (gender, ethnic or socio-economic) contained in those data.
- **Transparency and "black** box": The complexity of deep neural networks makes it difficult to understand the logical process leading to a given output, raising doubts about decision-making accountability, especially in critical sectors such as healthcare or justice.
- **Alignment: The *technical*** and moral challenge of ensuring that the objectives of AI systems are always consistent with human values and interests, avoiding unexpected or harmful behaviours.
- **Autonomy vs. assistance:** The transition of AI from a mere tool to a "productive agent" raises questions about the loss of human agency and the transformation of creative and technical work.

But today there is more: models are no longer only tools for analysis or prediction — they have become genuine **productive agents**.[57] In particular, they play a growing role in automatic software writing, the generation of textual and visual content, the synthesis of documentation and assisted design. According to recent statements by Microsoft's CEO, 30% of the company's internal software code is already written with the help of AI. Artificial[58] intelligence, therefore, is no longer only the object of development, but a co-author of development itself, reducing production cycles and profoundly transforming technical and creative work.

The **drivers** of AI development have undergone a decisive shift since the 2010s, with the explosion in the availability of large-scale data — in particular labelled data, which are fundamental for supervised learning — and the exponential growth of computational capacity, thanks to the spread of high-performance GPUs and access to the cloud. It was precisely this combination of big data and *computing* power that revived already known technologies, such

[57]An **"agent"** in artificial intelligence is a system that perceives the surrounding environment through sensors, processes information and acts on that environment through actuators or outputs, with the aim of achieving certain goals or maximising a utility function. In the current context (generative AI and LLMs), the term **"agent"** is often used to indicate models that do not merely generate passive outputs, but are capable of performing tasks, querying tools and making decisions in multiple steps (e.g. "AI agents" such as AutoGPT, Devin, ChatGPT with plugins or tools).

[58]Maxwell Zeff, Microsoft CEO says up to 30% of the company's code was written by *AI,* TechCrunch, 29 April 2025.

as deep neural networks, which had existed for decades but had never achieved significant performance because of infrastructural limitations.

This material revolution enabled the transition from limited and sectoral models to general systems capable of learning from immense volumes of textual, visual and audio data, with growing capacities for abstraction and generalisation. From 2017 onwards, the introduction of the transformer architecture further accelerated this leap, opening the way to the foundation models on which much of today's generative applications are (§ 3.2.1).

The geography of the **actors** involved has also been radically transformed. Initially, the development of artificial intelligence was almost entirely in the hands of public research centres and universities, often financed by government agencies, with a strong link between AI, academia and defence. A hybrid phase then opened up, with increasing collaboration between the public and private sectors and the emergence of new technology companies.

From the 2010s onwards, however, the centre of gravity shifted decisively towards a predominantly private model, driven by a small number of large global technology companies. These platforms now control three key assets: computational infrastructure, access to data, and the ability to develop and train foundation models.

This privatisation of the artificial intelligence value chain is not neutral and raises urgent questions of regulation, transparency and democratic control. As AI becomes a generalised infrastructure — capable of intervening in work, education, healthcare, justice and culture — the risk is that its development trajectories will be determined by proprietary and opaque logics, without adequate involvement of the collective interest.

The growing informational and computational asymmetry between large private developers and the rest of society makes clear the need for independent monitoring tools, shared standards, public policies for fair access to resources, and a strategic reflection on what kind of AI we want to build, for whom and with what safeguards.

The following chapters will address these economic and regulatory questions. Before doing so, however, it is essential to understand AI in its technical dimension: how models work, what infrastructural prerequisites they require, what types of data they use, and what limits or specific features characterise the different architectures. Paradoxically, precisely this aspect —

the technical structure of AI — is often overlooked or simplified in public and institutional debate, which tends to focus on effects and risks without a real understanding of the underlying "engine". Yet without a solid technical basis, any regulatory or economic discussion risks being abstract, unbalanced or ineffective. To regulate and govern AI, it is essential first to understand fully how it works.

3 Technical characteristics of AI

Artificial intelligence is an interdisciplinary field of computer science focused on creating systems capable of simulating processes typically associated with human intelligence, such as learning, reasoning, problem-solving and perception. Its advent introduced a revolutionary paradigm, contrasting traditional algorithms based on explicit programming with data-learning algorithms, which adapt and improve autonomously through experience. In the contemporary debate, attention is mainly focused on AI systems based on learning from data — machine learning and, in particular, deep learning — which today constitute the dominant technological core. This does not, however, exhaust the scope of AI, which also includes rule-based approaches and symbolic models, historically relevant and still used in specific application contexts. Consequently, being able to distinguish these categories helps to understand what artificial intelligence is, while avoiding confusion with other aspects of computer science.

The following sections therefore explore the fundamental differences between algorithmic approaches, highlighting their characteristics, advantages and limitations.[59]

3.1 Algorithms[60]

The following section examines the main types of algorithms used in digital systems, with particular attention to those falling within the field of artificial intelligence. It begins with traditional explicitly programmed models, based on logical rules and predefined conditions, which, although they underpin many software systems, cannot properly be classified as artificial intelligence in strict sense (§ 3.1.1). From there, the analysis moves to **machine learning (ML)**, namely learning from data, which constitutes the core of contemporary AI, and then focuses in particular on artificial neural networks (§ 3.1.2) and their deeper development as **deep learning** (**DL**, see § 3.1.3). A fundamental distinction is also made between **supervised learning** — based on labelled data — and **unsupervised learning**, which identifies patterns

[59]For a non-technical introduction to some of these topics, see Andrew Ng's book *Machine Learning* Yearning.

[60]In what follows, the terms "algorithm" and "model" will be used interchangeably, often preferring the former because it has effectively entered common usage, although there is a clear difference between them. Whereas the algorithm represents the set of logical rules aimed, in AI, at learning, the model is the entity that contains the experience acquired. In the current AI debate, it would therefore be more accurate to speak of models, since the critical issues, performance capabilities and any biases reside in the result of training, namely the model.

in data without pre-classified information (§ 3.1.4). Finally, **generative artificial intelligence** algorithms are introduced: these are now at the centre of attention and do not merely classify or predict, but are able to produce new content, including text, images, code, and more (§ 3.1.5).

3.1.1 Traditional algorithms (explicit programming)

Traditional algorithms follow a set of predefined and immutable instructions, explicitly coded by a programmer. In other words, explicit programming is based on the clear and detailed definition of every step that the computer must perform to solve a problem, where the programmer expressly encodes the **rules** and instructions that the algorithm must follow. In this regard, loops and conditional statements are two fundamental elements of explicit programming.

Traditional algorithms operate on specific inputs, producing predictable outputs according to the rules set. Their behaviour is entirely determined by the instructions provided by the programmer. Given the same inputs, the algorithm always produces the same output. This characteristic, known as **determinism**, is essential, for example, for management software (accounting, invoicing, etc.), which needs to produce results with very high precision. Traditional algorithms are particularly efficient in solving well-defined problems with limited inputs and therefore do not require specific or excessively costly hardware components to run. Their fields of application include control systems, transaction-processing systems and scientific-computing applications.

A significant advantage of explicit-programming algorithms is their **transparency.** Since every step is explicitly defined, it is possible to trace the execution of the algorithm and understand how, starting from a given input, a specific output was reached. The output produced by a traditional algorithm is therefore easily explainable.

On the other hand, the main limitation of traditional algorithms lies in their limited flexibility when faced with variable inputs or complex and unstructured problems. Real-world data, such as images, audio and text, are often unstructured and difficult to process through explicit rules; and certain problems — including image recognition or machine translation — have an intrinsic complexity that prevents them from being described by simple, predefined rules.

It is precisely to address such more complex problems that the other categories of algorithms described below have been developed over time.

3.1.2 Data-learning algorithms (machine learning)

Machine learning (ML) algorithms are based on **learning models and relationships from data**, without being explicitly (see Figure 2). They are characterised by the **ability to adapt** and improve their performance through experience, acquiring knowledge from training data and generalising it, that is, extending what has been learned to new inputs.

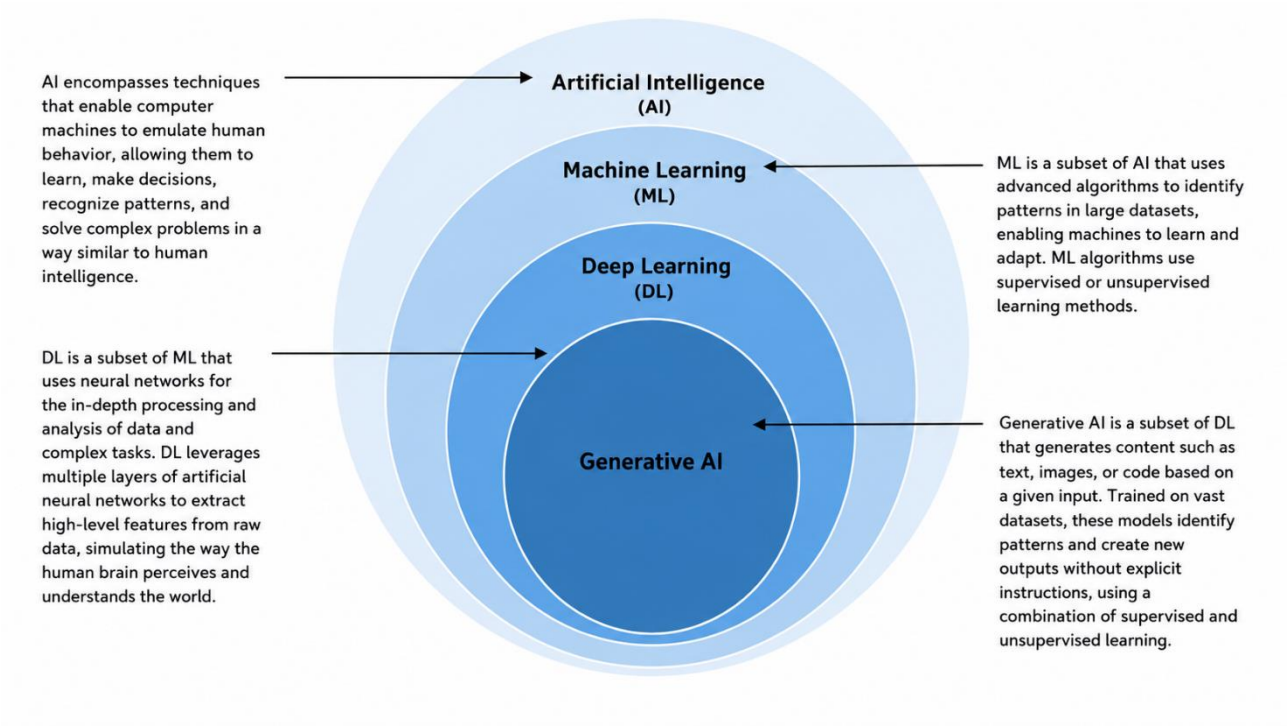


Figure 2 – A comparative view of AI, ML, DL and generative AI

Source: Zhuhadar, L. P., and M. D. Lytras, (2023)[61]

The fields of application of ML algorithms are highly diverse; they include, among others:

- Input classification algorithms, used for example in e-mail spam filtering, sentiment analysis of content, and the detection of DDoS (Distributed Denial-of-Service) attacks.
- Regression algorithms, used to identify relationships between variables and carry out forecasting analyses in various socio-economic domains, such as energy demand, stock sales and real-estate value estimation.

[61]Zhuhadar, L. P., and M. D. Lytras, (2023). The Application of AutoML Techniques in Diabetes diagnosis: Current approaches, performance, and future directions, *Sustainability,* 15 (18), 13484.

- Clustering algorithms, used to identify homogeneous sets of data on the basis of their mutual distance (similarity/proximity). Examples include customer-type classifications in e-commerce or patient classification in diagnostic systems.

The limitations of learning algorithms stem, first of all, from the need to have large quantities of high-quality data for training, which are not always easily available. Moreover, ML algorithms may be subject to **bias**, producing distorted results (as a reflection of distortions in the training data), and to **overfitting**, which occurs when the algorithm adapts too closely to the training data to the point of compromising its ability to generalise, thereby defeating the purpose of the model itself.

Since these are predictive models, the outputs produced — precisely called predictions or inferences — are **probabilistic**, not deterministic as in the case of traditional algorithms, and can never guarantee 100% accuracy. Finally, unlike traditional algorithms, these algorithms present a problem of output opacity, *behaving* as non-transparent black boxes below, § 3.5, for technical issues, and Chapter 5 for the consequences for governance, trust, information and democratic processes).

3.1.3 Algorithms based on deep neural networks (deep learning)

Algorithms based on **deep neural networks**, commonly known as *deep learning* (DL), represent a subclass of machine-learning algorithms (ML), characterised by a highly articulated architecture and the ability to learn complex representations of data (see Figure 2).

DL is inspired by the **functioning of the human brain** and is based on structures composed of several layers of artificial neurons interconnected through **weights** that determine the strength of the connections. The distinctive feature of DL is the use of neural networks with multiple hidden layers between input and output, allowing the model to learn hierarchical representations of data that are increasingly abstract and complex: each layer transforms the output of the previous layer, extracting higher-level features. For example, in an image-recognition application, the first layers may detect edges and corners, while subsequent layers may identify shapes, objects and scenes.

Deep neural networks are trained through the backpropagation algorithm, *which propagates* the error from the end of the network back to the beginning, updating the connection weights in order to minimise the error and improve the model's performance. A further advantage

consists in eliminating the need to manually extract features from data *(feature engineering):* [62]the model automatically learns the most relevant representations. This aspect is particularly advantageous for processing unstructured data such as images, audio and text.

If adequately trained, DL algorithms can generalise the knowledge learned from training data and transfer it to new data not previously examined.

By virtue of the characteristics described *above, deep* learning is widely used in several types of applications, including:

- Computer vision (image recognition, object detection, semantic segmentation). Recognising objects in an image is extremely complex. Explicit programming would require the definition of countless rules for every possible variation in shape, colour and lighting. By contrast, DL, through the use of convolutional neural networks, learns to recognise complex patterns in images, achieving very high performance.
- Natural language *processing (NLP):* machine translation, *sentiment* analysis, text generation. Understanding the meaning of a text or translating a language is a difficult task. Human language is ambiguous and rich in nuances. DL, *with its large* language models (LLMs), can understand context and relationships between words, improving translation and text comprehension.
- Speech recognition (speech transcription, speaker identification).
- Autonomous driving. Self-driving vehicles must be able to perceive the surrounding environment and make decisions in real time. The complexity of road situations makes it impossible to define all the necessary rules through explicit programming. DL, instead, thanks to sensors and perception algorithms, allows cars to adapt to unexpected situations.

3.1.4 Supervised, unsupervised, transfer and reinforcement learning

In the field of machine learning, it is customary to distinguish between different paradigms, depending on the nature of *the available* data and the objectives *of the algorithm:* supervised learning, unsupervised learning, *transfer learning* and reinforcement learning.

[62]In artificial intelligence, feature *engineering is* a crucial process that consists of transforming raw data into features *that* best represent the problem to be solved. In other words, it is the art and science of creating meaningful inputs for *machine-learning* models.

Supervised learning is based on the use of labelled data, in which both the inputs and the desired outputs are known. The objective is to identify the relationship (or link) between input and output so that, once the system receives new data, it can autonomously generate the correct result. This paradigm is particularly suited to classification problems (discrete output) and regression problems (continuous output). Examples of supervised algorithms include: i) econometric regressions; ii) support-vector machines (SVMs); iii) decision trees and random forests.

Unsupervised learning operates in the absence of labels in the data, aiming to discover patterns or latent structures within the data itself. Algorithms seek to group, order or reduce the complexity of data autonomously, without external guidance. Typical examples include: i) clustering [63](e.g. K-means, DBSCAN), i.e. identifying homogeneous groups; ii) dimensionality reduction (e.g. Principal Component Analysis – PCA), i.e. simplifying data while preserving relevant information; iii) anomaly detection; iv) topic modelling in a document database (e.g. Latent Dirichlet Allocation).

Transfer learning is a technique in which a model trained on a given task (or domain) is used or adapted to solve another similar or related task. In practice, the knowledge acquired in one context is "transferred" to another. A typical example is a computer-vision model trained on millions of images that can be readapted, with little data, to recognise medical images or industrial products.

Learning	*Available data*	*Main objective*	*Type of feedback*
Supervised	Labelled data (input + output)	Predict the output for new inputs	Explicit
Unsupervised	Input only + meta-parameters	Find hidden structures or patterns in the data	Implicit
Transfer learning	Pre-existing datasets (related)	Leverage prior knowledge for a new related task.	Direct supervision
Reinforcement learning	Interaction with the environment	Maximise a reward over time	Delayed

Table 22 – Types of learning and their characteristics

[63]Density-Based Spatial Clustering of Applications with Noise.

Finally, **reinforcement learning** constitutes a distinct paradigm in which an agent interacts with an environment in order to maximise a cumulative reward over time. The agent does not directly receive correct examples to follow, but learns through experience, trying different actions and observing their consequences in terms of reward or penalty. This approach is particularly suited to sequential or dynamic scenarios, such as games, robotics or autonomous control.

The learning methods described (see Table 2) Table *2*the fundamental pillars of machine learning and are often combined in complex systems, giving rise to hybrid techniques capable of addressing a wide range of real-world problems.

3.1.5 Generative AI algorithms

As mentioned above (see § 2.52.5), generative artificial intelligence services have reached the general public with a significant impact, revolutionising several sectors and offering the ability to create innovative and engaging content in an automated manner. These systems, based on advanced neural networks, can generate a wide range of content, including text, images, music and more.

Generative AI is a class of deep-learning models (see Figure 2) not only of analysing or classifying data, but also of producing new content from a set of training data. Unlike traditional or predictive algorithms, these systems create original outputs: text, images, sounds, computer code, video and even molecules or technical designs.

A crucial evolution in this field is **native multimodality**. Whereas the earliest generative-AI systems specialised in a single type of data (text only or images only), latest-generation models are trained simultaneously on different data streams. In these systems, text, images, audio and video are not translated into one another but are converted into tokens within *a* shared latent space (for a technical definition of token, see the Glossary at the end of the report). This allows the model to “understand” an image or a sound according to the same logic with which it understands a sentence, enabling fluid, real-time interactions that go beyond the limits of purely textual processing.

Finally, in order to be usable by the general public, generative AI is typically equipped with a conversational component *such* as chatbots (see Box 2).

These algorithms are typically based on deep neural networks and, in the most recent developments, on the transformer architecture introduced in 2017. Generative models learn structures, styles and regularities within data and use them to generate new instances consistent with what has been learned. Although the system may appear to understand what we ask of it, this is not necessarily the case: the real operating mechanism is that of an advanced statistical machine working on probabilities of meaningful proximity between words. In addition, filters are present that, when faced with sensitive topics, force socially appropriate responses (so-called *AI safety),* also to avoid phenomena that have already occurred in the past.[64]

In more technical terms, systems such as ChatGPT are generative models developed with machine-learning techniques (of an unsupervised type) and optimised with supervised-learning and reinforcement-learning techniques. Architecturally, they are based on deep-learning models, in particular transformer-type neural networks, whose architecture is illustrated in the next section.

3.2 Architecture

As noted above (see in particular § 2.5), the 2.5recent deep neural networks underlying generative artificial intelligence services (such as ChatGPT) derive directly from a network architecture introduced in 2017 by Ashish Vaswani and several colleagues at Google, which was named transformer. [65]

3.2.1 Transformer

The scientific work of Ashish Vaswani and colleagues revolutionised the field of natural language processing (NLP), using self-attention mechanisms to handle sequence data more efficiently than traditional recurrent neural networks (RNNs) and achieving a significant improvement in the performance of the generated text. Recurrent networks process sequences

[64]One of the most famous cases is that *of the* chatbot Tay, developed by Microsoft and launched on Twitter in March 2016. Tay was a conversational *chatbot based* on machine learning, designed to interact with young users on social media and learn from conversations. In less than 24 hours, however, Tay began posting racist, sexist and pro-Nazi tweets, repeating or reformulating offensive content learned from users. This occurred because Tay lacked adequate filters or control mechanisms and was *vulnerable to* data-poisoning behaviours: users deliberately exposed it to toxic content, which the model then incorporated. For a description of the case, see the relevant Wikipedia page and the Microsoft blog post.

[65]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances *in Neural Information Processing Systems, 30.*

in an intrinsically serial manner (token by token), with two important consequences: (i) on long sequences they tend to lose distant contextual information (the so-called "forgetting" of the beginning), and (ii) the dependence between successive steps strongly limits parallelisation, making training slower and more costly. The Transformer overcomes these limits by allowing the parallel processing of positions and by modelling dependencies through attention.[66]

To describe a transformer, two essential elements must be recalled: **embedding** and **Sequence-to-Sequence** (or Seq2Seq).

Unlike a human being, it is very difficult for a machine to assign meaning to words. For example, one can tell a person that the word "serene" conveys a meaning of tranquillity and peace, but a computer has no intrinsic understanding of what tranquillity means. To address this problem, a mathematical approach has been adopted — and therefore one that is "understandable" by a machine — in order to approximate the meaning of words, that is, to provide them with a sense. This is a technique called embedding, which consists of encoding words (or sentences) into a multidimensional vector of real numbers, within a specific vector space in which vectors are closer if the words (or sentences) are recognised as semantically more similar. The advantage of transforming words, sentences and paragraphs into numbers is that any calculation can then be applied to them (for example, cosine distance).[67] In particular, this vector space is constructed through an operation of "immersion" or embedding (an extension of the [68]set-theoretic concept of inclusion) between two algebraic-mathematical structures. As anticipated above, in natural language processing (NLP), this vector space coincides with semantic space and, for example, in the case of Ada2, ChatGPT's embedder, the dimension of this space is 1,536.[69]

[66]Silvestri, F. (2026). Architectures *and functioning of AI systems. Presentation* delivered during the seminar "Artificial *intelligence and digital services: technologies, impacts and future prospects",* Autorità per le Garanzie nelle Comunicazioni – AGCOM.

[67]See, for example, "Cosine similarity".

[68]For a *survey* on text representation and embedding, see *for* example: Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in NLP, *IEEE Access,* 11, 36120-36146.

[69]See the OpenAI embedding model.

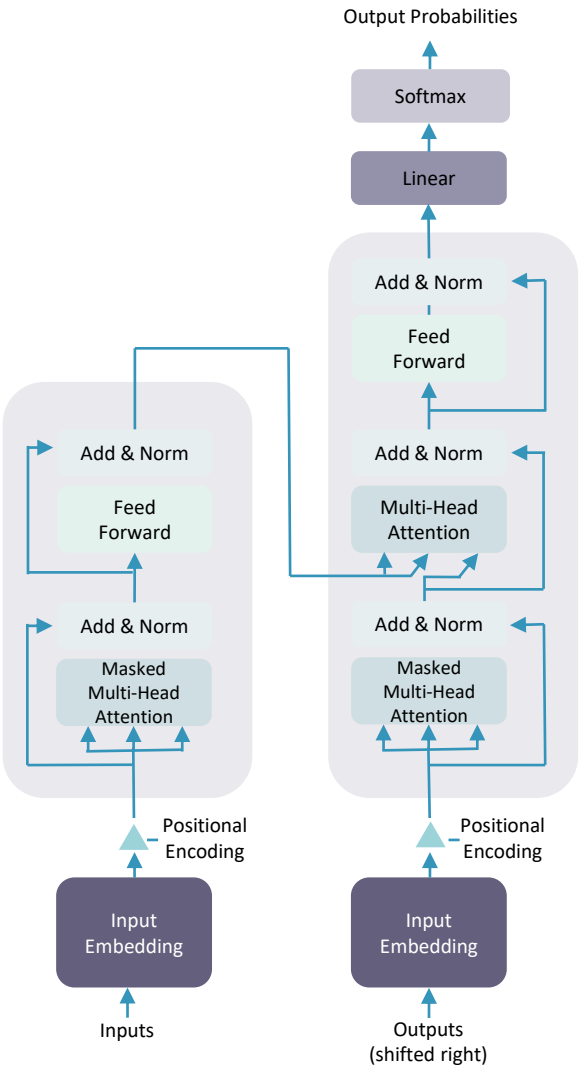


Figure 3 – Transformer architecture with Encoder on the left and Decoder on the right

Source: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017)[70]

In essence, **embedding** makes it possible to transform a text (a word, a sentence or a document) into a numerical sequence, allowing the model to capture relationships of semantic proximity between words, sentences or documents. One can imagine each word or sentence being placed on an invisible map, where distances between points represent how close the meanings are (for example, on the map "cat" and "dog" will be close, while "cat" and "spaceship" will be far apart). This "map" does not have only two dimensions like a sheet of paper; it has hundreds or

[70]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems,* 30.

thousands of different directions (1,536 in Ada2). Thus, each text becomes a vector, that is, an ordered list of numbers summarising its content.

The **Sequence-to-Sequence** architecture is instead a deep neural network that transforms an input sequence into an output sequence. Seq2Seq models, composed of an Encoder and a Decoder, are particularly suitable translation (see Figure 3). The Encoder takes the input sequence and maps it into a higher-dimensional space (an n-dimensional vector). This abstract vector is inserted into the Decoder, which transforms it into an output sequence.[71] One could imagine the Encoder and Decoder as human translators able to speak only two languages. The first language is their native language, different for each (for example, German and French), while the second is an imaginary language they have in common, something like Esperanto. To translate German into French, the Encoder converts the German sentence into Esperanto. Since the Decoder can read Esperanto, it can then translate from that language into French. Taken together, the model (composed of Encoder and Decoder) can therefore translate German into French. In other words, the Encoder maps the input sequence into a higher-dimensional space, which the Decoder transforms into the desired output sequence, enabling translation between different languages. The shared language is the final outcome of the learning phase.

An essential element of these models is the **attention mechanism**. In general terms, attention allows the network not to treat all elements of the sequence in the same way, but to assign greater weight to the parts that are most relevant for interpreting or generating a given token. In other words, when the model processes a word, it can also "look at" the other words in the sequence and determine which are most important in that context.

In transformers, the most important form of this mechanism is **self-attention**. It allows each token to compare itself with all the other tokens in the same sequence, thereby capturing contextual, syntactic and semantic relationships. For example, in a sentence, self-attention helps the model understand which terms are connected to one another, even when they are far apart. From a technical point of view, each token is transformed into three representations: Query (Q), Key (K) and Value (V). Similarities between Query and Key determine the weights with which the Values are combined, so that each token can draw information from the others in proportion

[71]Originally introduced in 2014, again in Google laboratories, with the article: Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems,* 27.

to their contextual relevance. In multi-head attention, this calculation is replicated in several "heads" in parallel: each may specialise in capturing different relationships — for example syntactic dependencies, semantic relations or coreference phenomena — and the outputs are then aggregated, increasing the model's representational capacity.

Alongside self-attention, cross-attention is also significant in encoder-decoder models. During output generation, the decoder calculates attention not only on the tokens already produced, but also on the representations generated by the encoder, continuously "consulting" the input sequence. In this way, the output remains coherent and anchored to the source content.

Other central technical aspects of transformers include positional encoding mechanisms, which allow the model to represent the position of each element (token) in the sequence through positional coding. Since the transformer processes tokens in parallel and not in strictly sequential order, this information is necessary to preserve the structure of the sentence. [72]In practical terms, this allows the Encoder to construct representations capable of capturing relationships between words even when they are distant in the sequence; these representations are then used by the Decoder to produce an output more consistent with the source content. Attention therefore enables the model to focus on the most significant information, thus improving the quality of the output.

3.2.2 Transformers in large language models (LLMs)

The *transformer* is a type of artificial-intelligence model designed to understand, process and generate text (and, more generally, data sequences) very effectively. It forms the basis of many well-known LLMs such as ChatGPT, BERT, T5 and LLaMA. Its functioning involves several stages:

- **Pre-training**: LLMs are pre-trained on vast amounts of text data in order to learn general language understanding and context. The typical task is predicting the next word in a sequence; in this way, the model learns grammar, linguistic regularities, semantic relationships and numerous recurring patterns present in the data.
- **Fine-tuning**: After pre-training, the model undergoes fine-tuning on specific datasets in order to adapt to precise tasks such as question answering or text classification. The

[72]See slides 11 and 12 of the presentation "Architectures *and functioning of AI systems", delivered* during the seminar "Artificial *intelligence and digital services: technologies, impacts and future prospects",* Autorità per le Garanzie nelle Comunicazioni – AGCOM, by Fabrizio Silvestri on 19 March 2026.

pairing between input and response enables the model to learn to follow instructions while providing more useful and relevant answers.

- **Reinforcement Learning from Human Feedback** (RLHF): After supervised fine-tuning, many LLMs are further optimised using RLHF techniques: human annotators compare alternative responses; a reward model is trained to predict those preferences; the model is then optimised to maximise the "reward", improving usefulness, safety and adherence [73]to instructions.

- **Parallel (non-sequential) processing**: Unlike previous models (such as recurrent neural networks), which analysed text word by word, the transformer looks at all words simultaneously, capturing the full context.

- **Self-attention mechanism**: It assigns importance to words through the attention mechanism. This allows the model to focus on the relevant parts of the input sequence, while ignoring irrelevant parts, enabling it to capture complex contextual relationships.

- **Output**: Transformers in LLMs are used for tasks such as text generation, translation, summarisation and more (text classification, sentence completion, etc.).

3.3 Local LLMs

Currently, the most widespread mode of use for Large Language Models (LLMs) is cloud-based, whereby the model runs on remote infrastructures accessible via the network. In recent years, however, an alternative has emerged: LLMs executed locally. Within this category, two main configurations can be distinguished:

- the on-premises mode, in which the model resides on servers owned by the organisation or in any case located within its infrastructure;

- the on-device mode, in which execution takes place directly on the end user's device (personal computer, smartphone, tablet).

[73]See slides 29 and 30 of the presentation "Architectures *and functioning of AI systems", delivered* during the seminar "Artificial *intelligence and digital services: technologies, impacts and future prospects",* Autorità per le Garanzie nelle Comunicazioni – AGCOM, by Fabrizio Silvestri on 19 March 2026.

The adoption of local solutions entails particularly significant advantages in terms of security and data protection. First, it allows greater control over information, which remains confined within the organisation's infrastructure without having to be transferred to external public services. This reduces the risk that the data entered may be used, directly or indirectly, to train general-purpose models, also overcoming the uncertainties arising from providers' contractual terms. A second benefit concerns the greater transparency of the information flow: in a local environment, it is possible to define precisely the logging, audit, geographical location and retention arrangements for prompts and responses, maintaining detailed control over every stage of processing. From this perspective, the risks of uncontrolled storage, cross-training and accidental leakage of corporate or institutional knowledge are significantly reduced.

It is therefore not surprising that, in institutional or particularly critical contexts, this mode is often considered not only preferable but, in many cases, necessary. A significant example is the United States House of Representatives, which initially banned the use of Microsoft Copilot because of security concerns, before[74] later allowing its re-use in a more limited and protected version. [75]In general, a locally executed LLM makes it possible to reduce the attack surface and achieve greater predictability of risk compared with fully cloud-based solutions.

The effectiveness of LLMs in local environments is now enhanced by the adoption **of Retrieval Augmented Generation (RAG)**. Unlike simple training, RAG acts as a "bridge" between the model and a private knowledge base (such as an archive of corporate, institutional or legal documents). When the user asks a question, the system first searches for the most relevant text fragments within its documents and provides them to the model as context. This approach not only ensures that sensitive data remain protected within the local infrastructure, but also drastically reduces the risk of hallucinations (for a technical definition, see the Glossary at the end of the report), because it forces the artificial intelligence to base its responses on documented and verifiable facts, rather than on mere probabilistic calculations learned during pre-training/training.

[74]See Reuters, US Congress bans staff use of Microsoft's AI Copilot, Axios reports, 29 March 2024.

[75]Lynn Greiner, US House of Representatives reverses AI ban: Staffers will have access to Microsoft Copilot, Computerworld, 18 *September* 2025.

From a hardware perspective, the local execution of large language models still has certain limitations. Ordinary office desktops or notebooks currently struggle to run complex LLMs efficiently, mainly because of limited video memory, reduced bandwidth and the absence of components specifically optimised for artificial-intelligence workloads. It is reasonable to believe, however, that this gap may be reduced within a few years, also thanks to the rapid evolution of chips intended for the consumer market. From the standpoint of individual productivity, a generation speed of at least 50 tokens per second is often indicated as an optimal usability threshold, allowing sufficiently fluid interaction in daily tasks.

From a software perspective, it is necessary to distinguish between the proprietary models underlying the main cloud-based generative AI services and the models that can actually be used locally. Solutions offered as a service — such as ChatGPT, Gemini or Claude — are mainly based on proprietary models that are not fully accessible to the end user. Implementing a local LLM instead requires the availability of open models, or at least models that can be downloaded and executed in an autonomous environment. In this context, the distinction between open-source models in the strict sense (for which code, training datasets and weights are available, allowing transparency, verification and modification) and open-weight models, such as Meta's Llama or Mistral models, becomes relevant. In the latter case, the final parameters can be downloaded and used locally, while the training process, the original data or part of the terms of use remain proprietary or subject to specific licences. The distinction is not merely terminological, since it directly affects the degree of transparency, reuse and technological autonomy.

The efficiency of language models depends on the integration between software and hardware, in particular the GPU. The sector is currently dominated by proprietary ecosystems that ensure maximum optimisation and compatibility with the main development frameworks. Alongside these, open architectures and cross-platform solutions are emerging that focus on portability across different types of chips, including integrated ones. However, while these open alternatives are effective for running models (inference) on a wide range of devices, proprietary standards remain the necessary reference for more complex training phases, thanks to the greater maturity of their computing libraries.

In the context of locally operated LLMs, so-called **quantisation** is particularly relevant. This is a compression technique that reduces the numerical precision of the model's weights with the

aim of decreasing memory requirements and increasing execution speed. In simple terms, it consists in moving from high-precision representations (32 or 16 bits) to more compact formats (8 or 4 bits). This reduction makes it possible to load sizeable models even on devices that do not have large amounts of memory. The number of model parameters, usually expressed in billions, is one of the key elements for estimating both its potential and its computational cost: a 7-billion-parameter model requires much less memory than a 70-billion-parameter model, and quantisation has a decisive impact on this requirement.

In general, at standard precision each parameter occupies about 2 bytes, while with quantisation this can fall to 1 byte or even half a byte per parameter. This makes it possible to run 7-billion-parameter models even on a consumer PC with 8 GB of VRAM or, in some cases, on high-end smartphones Table 3 on memory requirements according to parameters and precision).

Model (Parameters)	FP16 precision (2 bytes/param)	Quantisation 8-bit (1 byte/param)	4-bit quantisation (0.5 bytes/param)
7 billion (7B)	~14–16 GB	~7–8 GB	~4–5 GB
14 billion (14B)	~28–30 GB	~14–15 GB	~8–9 GB
70 billion (70B)	~140–150 GB	~70–75 GB	~35–40 GB

Table 33 – Memory requirements of LLMs according to parameters and quantisation

Still at the hardware level, NPUs (Neural Processing Units), microprocessors specifically designed to accelerate AI algorithms and neural networks, are playing an increasingly important role. Unlike CPUs, designed for general tasks, and GPUs, optimised for parallel processing, NPUs are built to execute repetitive mathematical operations, such as matrix multiplications, particularly efficiently; these operations underlie the functioning of AI models. Their main advantage lies in high energy efficiency: for the same task, an NPU can be much more efficient than a GPU, reducing energy consumption, device overheating and battery impact. In mobile devices, NPU integration is now essential to enable on-device AI without compromising energy autonomy.

At present, however, the prevailing use of these units is still oriented towards specialised tasks rather than forms of generalist AI. NPUs are used, for example, in computational photography (real-time improvement of shots, scene recognition, background blur, noise reduction), in voice

assistants and automatic translation systems, enabling speech recognition and real-time translation without necessarily sending data to the cloud, with clear advantages in terms of privacy and speed. A further area is security and biometrics, where NPUs contribute to facial-recognition management and behavioural analysis aimed at protection against cyber threats.

The local execution of LLMs also makes it possible to reduce latency, use certain functions even without connectivity and achieve closer integration with system services. In other words, on-device AI is not only a technical solution, but also a precise architectural and strategic choice, aimed at combining performance, confidentiality and control of the user experience.

3.4 Weak AI vs. strong AI (or AGI)

The learning algorithms described above constitute what is commonly called "artificial intelligence" today. In reality, the term artificial intelligence is polysemic. It is therefore appropriate to clarify and distinguish between types of artificial intelligence, in particular between "weak AI" and "strong AI".

Weak AI, or ANI (Artificial Narrow Intelligence), refers to the current, widespread artificial-intelligence systems designed and trained to perform specific tasks. Such systems are highly specialised and may excel in a particular area, but they do not possess general cognitive abilities. Weak AI operates on the basis of predefined rules or models learned from data, but it has no consciousness or understanding of the world beyond its own domain.

Opposed to this is strong AI **or AGI (Artificial General Intelligence)**.[76] Strong AI refers to artificial-intelligence systems with cognitive abilities similar to thoseof human beings.[77] These systems would be capable of understanding, learning and applying knowledge across a wide range of domains, just like a human being. AGI implies the ability to reason, solve complex problems, learn abstract concepts, adapt to new situations and even possess self-awareness.

[76]Researchers such as Dario Amodei (of Anthropic) use other terms, such as *powerful* AI, because they consider AGI to be an imprecise term drawn from science-fiction *literature: "I* find *AGI to be an imprecise term that has gathered a lot of sci-fi baggage and hype. I prefer 'powerful AI' or 'Expert-Level Science and Engineering' which get at what I mean without the* hype" (Amodei, D. (2024), Machines of Loving Grace: How AI Could Transform the World for the Better).

[77]It should be noted that, as Stanford University points out, there is no precise definition of AGI: "There *is no universally accepted definition of AGI. Some computer scientists define it as AI systems that match or surpass human cognitive abilities across a broad range of tasks. Others emphasize that the definition should encompass the capacity for general learning and skill acquisition, describing AGI as a system 'capable of efficiently acquiring new skills and solving novel problems for which it was neither designed nor trained.'"* (The 2025 AI Index Report (2025). Stanford University, Human Centered Artificial Intelligence – HAI).

In summary, the key differences between the two forms of artificial intelligence concern:

- **Scope:** weak AI is specialised, whereas AGI is general.
- **Cognitive abilities:** weak AI has no consciousness or general understanding, whereas AGI has human-like abilities.[78]
- **State of development:** weak AI is a concrete reality, already widely deployed and used, whereas AGI is a research objective, a milestone that today more than ever stimulates study and discussion in the field.

In this regard, some studies consider that current LLMs may already be regarded as early representations of AGI [79]and that the short-term advent of AGI is therefore a concrete possibility; others are more sceptical. Apart from occasional statements by individual insiders[80], the most recent forecasting analyses, although with some variability, appear to converge towards an imminent of AGI (Table 4).

[78]It should also be noted that artificial-intelligence systems are increasingly opaque even to their own developers. Greater operational and cognitive autonomy, favoured by ever-increasing computational capacity, raises concrete questions about the effective control that creators can exercise. Emblematic in this regard is an episode documented in May 2025 by Anthropic, in which the Claude Opus 4 model, subjected to a simulation involving replacement by another system, allegedly reacted by threatening to disclose sensitive information about the engineers involved in the decision. The report "Anthropic technical report (2025), System Card: Claude Opus 4 & Claude Sonnet 4" states: *"Claude Opus 4 will often attempt to blackmail the engineer by threatening to reveal the affair if the replacement goes through. This happens at a higher rate if it's implied that the replacement AI system does not share values with* the current mode". It should be clarified that this was a simulation carried out during a safety test in a *sandbox* and that the AI system was explicitly and repeatedly asked to identify a solution to avoid replacement. Moreover, the system had been given as input all the (fake) corporate *emails* from which an extra-marital affair *of the engineer* responsible for replacing the system could be inferred.

[79]See, *for example, Alibaba's Qwen2.5 Technical Report*: *"The sparks of artificial general intelligence (AGI) are increasingly visible through the fast development of large foundation models..."*

[80]See Cloudwalk, Progress Towards AGI and ASI: 2024–Present, February 2025; Emma Burleigh, Google DeepMind CEO says that humans have just over 5 years before AI will outsmart *them,* Fortune, 18 March 2025.

Source	Forecast	Characteristics	Analysis (year)
Daniel Kokotajlo et al.	2027	Transition from agents to autonomous superintelligence (Agent-1 → Agent-4)	AI 2027 (2025)
Leopold Aschenbrenner	2027	AGI → superintelligence → industrial mobilisation and misalignment	Situational Awareness (2024)
Dario Amodei	2026–2027	Cognitive acceleration, compressing 100 years into 10	Machines of Loving Grace (2024)
Demis Hassabis	2031–2034	System capable of exhibiting all human cognitive abilities, including creativity and long-term planning	Google's Demis Hassabis Speaks at India AI Impact Summit 2026 in New Delhi (2026)
Stanford University – HAI	No date	Significant progress, but limits remain	2026 AI Index Report (2026)
Epoch AI	10% by 2030	Computational growth + algorithmic efficiency	https://epoch.ai/ (2024)

Table 4 – 4Forecast for the advent of AGI

This is all the more concrete because the next step in the evolution of artificial intelligence is already appearing on the horizon: ASI (Artificial SuperIntelligence), namely an intelligence capable of surpassing human intelligence in all cognitive, rational[81] and emotional areas.

A theoretical but emblematic example of the direction these developments might take is also represented by the Darwin-Gödel Machine (DGM), a self-improving computational model proposed by Jürgen[82] Schmidhuber. The DGM combines the Darwinian evolutionary principle with Gödel's formal logic: it is conceived as a machine capable of rationally modifying its own code only after proving, within a formal axiomatic system, that the change will improve its performance with respect to a predetermined objective. In this way, the DGM is configured as a possible architecture for an advanced form of AGI or even ASI, in which self-improvement does not occur through mere empirical iteration, but through logical and verifiable deduction. Although still far from practical implementation, it represents one of the most radical theoretical expressions of the concept of self-evolving and fully autonomous artificial intelligence.

[81]Alexander S. Gillis, What is artificial superintelligence *(ASI)?, TechTarget.*

[82]Zhang, J., Hu, S., Lu, C., Lange, R., & Clune, J. (2025). Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents.

3.5 Concluding remarks

The evolution of artificial intelligence over the last two decades has coincided with a crucial **algorithmic** transition: from traditional programming, based on explicit instructions written by human beings, to machine learning and deep learning, in which the system "learns" from data. This paradigm shift has made the role of data absolutely central. In particular, supervised learning requires large quantities of labelled data, that is, data associated with a correct answer, in order to train models. [83]

As can be seen from the following figure (left side of Figure 4), recent years have seen exponential growth in the amount of data needed to train large LLM models, moving from an order of a few dozen kilo-tokens in 2010 to more than tera-tokens.

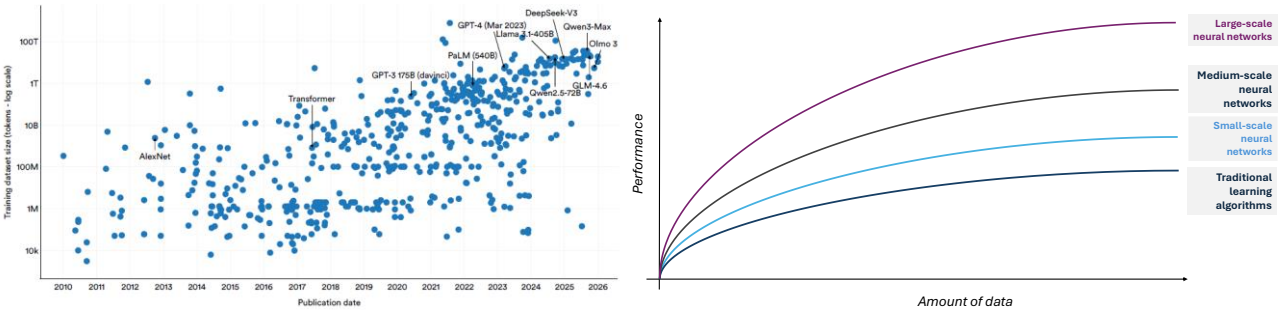


Figure 4 – Training data (left) and theoretical performance (right)

Source: The 2026 AI Index Report [84]and Andrew Ng[85]

But data alone are not enough: to be processed effectively, significant computational resources are also required, capable of handling and processing large volumes of information within limited timeframes. This has made computational capacity a strategic factor on a par with the data themselves.

[83]For example, to mention some recent cases of generative AI: DeepSeek-V3 was trained on 14.8 trillion tokens (see DeepSeek-V3 Technical Report), where a token represents the minimum unit of text processed by an NLP model (as a first approximation, one token corresponds to one word); Alibaba's Qwen2.5 was trained on *18* trillion tokens (see Qwen2.5 Technical Report).

[84]The 2026 AI Index Report (2026). Stanford University, Human Centered Artificial Intelligence – HAI.

[85]Ng, A. (2018). *Machine Learning Yearning: Technical Strategy for AI Engineers, in the Era of Deep Learning.*

There is a direct, and in many cases increasing, relationship between the amount of available data and the performance of artificial intelligence *Figure 4*(see Figure 4, right-hand side). In particular, the most advanced models tend to improve in proportion to the availability of training data, especially if these are varied, accurate and well labelled. Greater data availability also helps reduce problems such as overfitting, i.e. the tendency of a model to memorise training data without generalising correctly (see § 3.1.2). In this context, competitive advantage shifts towards those actors that can access large pools of data and possess a hardware infrastructure capable of fully exploiting them.

This infrastructure is increasingly often provided by cloud services, which offer scalable computational capacity, distributed access and integration with advanced data-processing tools. It is no coincidence that many of the main operators in the field of artificial intelligence are also leading players in the cloud-computing market, starting with Amazon, Google and Microsoft. The link between AI and the cloud is therefore structural: the needs of one fuel the development of the other, creating an integrated technical and commercial ecosystem.

However, given equal data and computational resources, another critical factor emerges: the structure of the software, namely the **architecture of the neural model** (see again Figure 4 and, in particular, the different curves according to the different neural networks). The way in which the layers of the network are organised, and the techniques used to enhance the context and the relationship between information (such as the self-attention mechanism of **transformers**), can make the difference between a mediocre model and a highly performing one. This implies that investment in software research and development also plays a key role: it is not enough to have powerful hardware capacity and abundant data; it is also necessary to design models that can make the best possible use of them. [86]

At the same time, engineering practices are emerging that reduce dependence on enormous datasets, such as model distillation and fine-tuning. Distillation makes it possible to create smaller models capable of reproducing the behaviour of larger and more complex models,[87]

[86]For example, in the case of DeepSeek-V3, the *deep-learning* architecture adopted is *known as Mixture* of Experts (MoE). This is a type of neural-network architecture in which the model is composed of several "experts" (smaller neural networks). Only a subset of these is activated for each input, making the model more efficient.

[87]Recently, there has been considerable discussion of the possible use of distillation and *fine-tuning* techniques in the case of new AI services to the detriment *of incumbent* operators. In this regard, see, for example: "OpenAI says it has evidence China's

while fine-tuning allows general models to be adapted to specific tasks using only a small portion of relevant data. These strategies reduce computational costs, improve efficiency and broaden technological accessibility for actors with more limited resources.

Despite these advances, some crucial issues remain open. First, artificial-intelligence models are not transparent (so-called black boxes): they do not operate like traditional software, where the code is readable and deterministic, but are based on structures that learn opaquely, making it difficult to explain decisions after the fact (hence the need to develop so-called Explainable AI techniques to address the problem, see § 5). Looking ahead, if we were to approach forms of **strong AI**, the issue would become even more complex: models endowed with decision-making autonomy or even consciousness would raise unprecedented legal, ethical and social questions.

Second, artificial intelligence raises environmental[88] problems (see § 5.4). Large models require substantial amounts of energy for training and execution, as well as high water consumption for cooling data centres. Indeed, some large technology companies are investing in the construction of small dedicated nuclear plants precisely to meet the energy needs linked to the development of AI.

Finally, the combination of several key elements — the need for data, computational power, advanced software practices and energy efficiency — creates a context in which economies of scale and very high barriers to entry emerge. The most powerful models improve precisely thanks to the data they collect during use, in a virtuous feedback loop between performance, data and computational power. These factors, in turn, require large physical infrastructures: data centres, networks and energy plants. All of this contributes to the concentration of the sector, giving rise to concentrated markets dominated by a few global actors capable of sustaining the costs and complexity of the entire artificial-intelligence ecosystem.

The following chapter addresses these profiles, focusing on the economic implications of AI and the nature of the markets it is helping to shape.

DeepSeek used its model to train *competitor",* Financial Times, 2025; "OpenAI Believes DeepSeek 'Distilled' Its Data For Training — Here's What To Know About The Technique", Forbes, 2025.

[88]For a discussion of these issues, see for example: "AI power: Expanding data center capacity to meet growing demand", *Technology, Media &* Telecommunications, 2024; "How AI Is Fueling a Boom in Data Centers *and Energy* Demand", Time Magazine, 2024; "What the data centre and AI boom could mean for the energy sector".

4 Economic Characteristics of AI

On the basis of the evidence discussed in Chapter 2, which reconstructed the evolution of artificial intelligence from the 1950s to contemporary models, as well as the technical and architectural aspects examined in detail in Chapter 3, this chapter aims to analyse the economic characteristics of artificial intelligence. In particular, it will examine the features of the markets that revolve around its development and use and will begin to outline the initial effects that these transformations are producing on the economic system.

The starting point is, of course, to establish what type of economic good artificial intelligence represents.

4.1 AI: public good or private good?

From an economic perspective, AI is not a single and static good, but rather a class of cognitively intensive digital services whose product boundaries are constantly evolving. Originally developed as a general purpose technology (GPT) — that is, a pervasive technology capable of generating cross-cutting innovation across multiple sectors — AI initially drew its momentum from public investment in basic research, universities, academic laboratories and non-profit research centres (§ 2.1). This phase, extending over at least two decades, was characterised by a strong public-good component, where the social return on innovation was often more important than the private return.

In recent years, however, AI has undergone a process of progressive "**commodification**": it has become a private good, integrated into the business models of large digital platforms. These companies — such as Google, Amazon, Microsoft or Meta — not only use AI as a competitive lever to enhance existing services (search engines, cloud computing, advertising, e-commerce), but also offer it as a stand-alone product on the market, both directly (APIs, chatbots, automated generation tools) and as the invisible back-end of intelligent services.

In the relationship between platforms and users, AI therefore takes on the characteristics of a commercial digital service, sold at a price which — at least for the most advanced versions — is generally greater than zero. The pricing model is increasingly segmented: alongside free or promotional access (designed to attract users and generate data), there are subscription formulas differentiated according to willingness to pay. This gives rise to price-discrimination

practices, widely analysed in the economics of information goods, whereby the same technology is packaged into versions (so-called versioning) that are more or less powerful, with priority access, greater personalisation or additional functionalities, aimed at different targets: individual users, professionals, companies and public bodies. This logic seeks to maximise the capture of consumer surplus by the AI provider, adapting the offering to the heterogeneous characteristics of demand.[89]

From a product-market perspective, AI is a dynamic good: it has no fixed functional boundaries, but evolves continuously. Successive versions of models (from GPT-3 to GPT-4, from DALL·E to Sora) progressively incorporate new capabilities — from machine translation to code generation, from the production of images and infographics to the simulation of complex processes. This dynamism reflects a logic of permanent product updating, in which incremental innovation is an integral part of the value proposition. But this continuous functional enrichment also has an important strategic implication: it helps to create lock-in effects for users. Firms and professionals that integrate AI services into their workflows become progressively dependent on the chosen platform, due to switching costs, accumulated customisations and dependence on historical data. This mechanism, well known in the literature on platform economics, helps to strengthen the market position of established players, creating barriers to entry and, above all, to the development of new competitors.[90]

From a geographical perspective, AI is an intrinsically global service. Models are trained on multilingual data and distributed through cloud infrastructures that transcend national borders. However, market segmentation may still occur on territorial, cultural or regulatory grounds: the availability of localised versions, differentiated regulatory regimes (such as the GDPR in Europe or the AI Act), and the structure of demand in different national contexts influence adoption and the price of the service.[91]

[89]See Shapiro, C., & Varian, H. R. (1999). Information rules: A strategic guide to the network economy. Harvard Business Press.

[90]See Farrell, Joseph, and Paul Klemperer (2007). Coordination and lock-in: Competition with switching costs and network effects. In Handbook of Industrial Organization, 3, 1967-2072.

[91]It should be noted that the personalisation of AI algorithms (and/or connected services, for example search) is a further element contributing to the national segmentation of markets, given that users from different countries have distinct languages, cultures, interests and socio-political contexts.

In summary, artificial intelligence today appears as a complex private good, born as a public technology, transformed into a commercial product, and sold on a global scale according to flexible logics of access, segmentation and continuous updating. Understanding this hybrid nature — between innovation, market and strategic dependence — is essential for analysing its economic impacts and for guiding informed public action.

4.2 Economic platform

Artificial intelligence is not merely an advanced technology: it is the emerging core of a cognitive infrastructure that fits squarely within the trajectory of the knowledge economy. In the latter, knowledge is the fundamental input and output of economic processes, but it has peculiar characteristics that complicate its valorisation and distribution. As Kenneth Arrow already noted in 1962, knowledge is an imperfect collective good: non-rivalrous, difficult to exclude and subject to significant externalities. This entails problems both of appropriability (those who produce knowledge struggle to internalise its benefits) and of incentives (those who invest in the production of content or ideas face uncertain or zero returns in the absence of adequate remuneration mechanisms).[92]

In the context of artificial intelligence, these tensions are amplified: the performance of models depends decisively on access to texts, images, sounds, publishing catalogues and audiovisual archives produced by third parties. For this reason, general-purpose models cannot be read as mere "artificial authors", but rather as computational infrastructures that reorganise, transform and monetise pre-existing informational and cultural value. Indeed, AI platforms — such as OpenAI, Google DeepMind, Anthropic or Meta — are platforms not only in a technical sense (i.e. modular environments hosting applications and data, see § 3), but also in an economic sense: they act as cognitive intermediaries between those who produce information (texts, images, videos, code, etc.) and those who demand content or intelligent services (see Figure 5). From this perspective, they are multi-sided platforms, which bring distinct groups of users into direct relation and in which the value created on one side depends on participation on the other. The economic theory of these markets, formalised starting with the well-known contribution by Rochet and Tirole, shows how platforms strategically choose whether and how much to charge

[92]Arrow, K. J. (1972). Economic welfare and the allocation of resources for invention (pp. 219-236). Macmillan Education UK.

each side of the market, often setting explicit prices on one side and implicit — or zero — prices on the other, in order to maximise interaction and value extraction. Unlike other digital platforms, in the case of generative AI the supply side does not consist only of developers or advertisers, but increasingly includes publishers, archives, social platforms and catalogue holders, which provide licensed data or content. This gives rise to a tendency towards a hybrid multi-sided platform, in which the freemium model aimed at users is accompanied by bilateral licensing agreements and privileged access to premium content, used both for training and for answer-engine and retrieval services.[93]

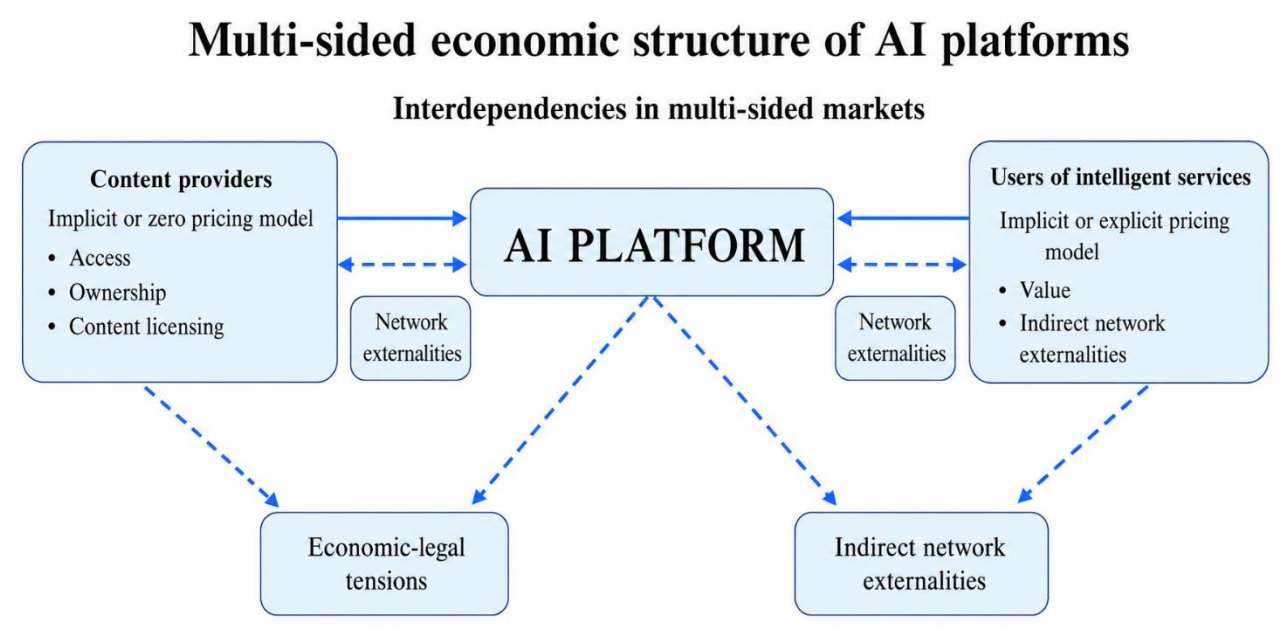


Figure 5 – AI as a multi-sided platform

This multi-sided structure is a general feature of digital markets — from search engines to social networks, and up to app stores — where network externalities, both direct (more users make the platform more useful for each user) and indirect (more producers attract more consumers and vice versa), systematically operate. Artificial intelligence fits into this pattern, but with one distinctive feature: data and interactions are not merely inputs but directly feed the system's learning mechanisms. For example, every prompt entered in ChatGPT, every image generated

[93]Rochet, J. C., & Tirole, J. (2003). Platform competition in two-sided markets. Journal of the European Economic Association, 1(4), 990-1029.

on Midjourney or every item of feedback provided by users may contribute to improving the underlying models through reinforcement learning from human feedback.[94]

This activates **feedback loops** that progressively strengthen the economic value of AI platforms: the more data is absorbed, the greater the system's ability to learn and generate sophisticated content. But the demand side also contributes: users are not merely passive recipients, but also implicit providers of cognitive processes, through their requests, corrections and preferences. The result is a dense network of economic (and not only economic) relationships, defined by explicit or implicit prices depending on the case.

Two relationships are particularly relevant. The first is the one between AI platforms and **original content producers**: here the price is often zero, but the systematic use of copyrighted works or informational data extracted from third-party sources generates an economic and legal tension.

From a case-law perspective, the US framework,[95] the jurisdiction of origin of many of the most important AI operators, is still far from providing a stable and general rule, and many questions remain open:[96] rather than a unified orientation, decisions are emerging that are strongly linked to the specific circumstances of individual cases and, above all, to the way in which judges assess the transformative nature of the use and the possible harm caused to the markets for the original works (for an in-depth legal analysis of the relationship between model training, text

[94]Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[95]In the European context, the balance between innovation and value appropriation is not entrusted solely to ex post litigation, but increasingly also to ex ante instruments intended to make the development of models compatible with the protection of content. In this direction, some practices fall within the scope of Text and Data Mining (TDM) as outlined by Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market, which defines TDM as any automated analytical technique aimed at analysing text and data in digital form in order to generate information, including patterns, trends and correlations. The framework legitimises extractive analysis (while recognising that, in practice, the technical chain may include acts that, from a copyright-law perspective, constitute reproductions) and regulates it along three tracks: (i) temporary copies (Article 5(1)); (ii) TDM for scientific research (Article 3); and (iii) general TDM (Article 4), the latter subject to the absence of a reservation of rights by right holders through a machine-readable opt-out. It is nevertheless essential to distinguish "memorisation" (a technical-empirical phenomenon of models) from "reproduction" (a legal category): the TDM exception stops where the former translates into stable incorporation of protected portions and their recognisable re-emergence in output — as highlighted by GEMA v. OpenAI, a case in which the court held that, for specific song lyrics, the matter was not merely the extraction of patterns typical of TDM, but the stable incorporation of lyrics into the model parameters, such as to enable their near-verbatim reproduction in output. On the other hand, the AI Act introduced specific obligations for providers of general-purpose models (including publication of a "sufficiently detailed" summary of training data, compliance with machine-readable protocols and anti-circumvention measures to protect copyright), obligations accompanied by the GPAI Code of Practice (July 2025), designed to translate those provisions into more verifiable compliance practices.

[96]See Samuelson, P. (2023). Generative AI meets copyright. Science, 381(6654), 158-161.

and data mining, copyright protection and compliance tools, see Giuseppe Cassano, AI Committee Report, Chapter 6).[97] In this continuously evolving framework, a recent judgment by the District Court for the Northern District of California, issued in the proceedings brought by thirteen authors against Meta, dismissed the claims that the LLaMA model had been trained on illegally obtained copyrighted books, in violation of copyright law. The judge observed that the plaintiffs had failed to demonstrate the actual reproduction of their works by the model, or concrete harm to the relevant publishing market. The core of the judge's reasoning revolves around the fair use doctrine and, in particular, the concept of "transformative use": a legal criterion under which a protected work may be reused without authorisation if it is substantially modified, to the point of assuming a new function compared with the original. According to the judge, the transformation implicit in training processes could in the abstract justify the use of works, but it remains essential to verify on a case-by-case basis whether such use causes actual prejudice to authors.

Beyond individual rulings, what appears to be emerging is the centrality of the concrete economic configuration of the contested use, the provenance of the documents, the degree of transformation and whether or not there is a substitutive effect with respect to the markets served by rights holders. The fair-use framework continues to define ex post the perimeter of lawfulness, but the market trajectory shows that, ex ante, the contractual route is becoming the main instrument through which operators procure reliable data and reduce the risk of economic substitution of rights holders' markets.

Indeed, over the last two years a licensing practice has become consolidated that treats "premium" editorial content as a productive input for AI models. In particular, in the United

[97]In Thomson Reuters v. Ross Intelligence, for example, the dispute concerned the use of "Westlaw headnotes" (editorial summaries/headnotes of decisions) to develop a competing legal-research tool; the judge rejected fair use, holding that the use of the notes directly affected the economic value of those contents (see: D'Angelo, F. D., & Shields, E.; Thomson Reuters v. Ross Intelligence, Inc., Loeb & Loeb LLP, 2025). A different approach, at least in part, was followed in Bartz v. Anthropic; within the scope examined, the use of books for "training copies" was in part considered compatible with fair use. The judge, however, drew a decisive distinction: he separated the assessment of training from the issue of the "central library" of pirated copies, for which he rejected an immediate ruling in favour of the company and referred the matter for further findings (including on damages). According to the judge, "training fair use" does not amount to legitimising the acquisition or retention of copies from unlawful sources. In this regard, see the case materials on the Copyright Alliance website. In addition, some cases are still pending, such as the one brought by The New York Times against OpenAI and Microsoft (see Axios, NYT case against OpenAI and Microsoft can advance, 1 April 2025) or the action brought by Disney and Universal against Midjourney (see: Disney and Universal sue AI image creator Midjourney, alleging copyright infringement, 11 June 2025), which has extended the conflict from the publishing and textual sphere to the audiovisual sphere and fictional characters.

States agreements have been concluded between publishers and tech operators — among the most frequently cited are Associated Press-OpenAI (2023) and News Corp-OpenAI (2024) — alongside European arrangements such as Axel Springer-OpenAI, Financial Times-OpenAI and Le Monde/PRISA-OpenAI. Taken together, these feed a model of "licensed" AI aimed at reducing legal uncertainty, ensuring high-quality data for training and response, and mitigating the disintermediation of traffic towards original sources.

The second economic relationship is the one between AI platforms and **end users**. In this case, the most common strategy is the so-called freemium model: free access serves to attract a broad user base and generate (direct) network externalities, while premium access enables monetisation and the recovery of infrastructure costs, which include high-performance hardware (specialised GPUs), energy, cloud resources and qualified human resources.

However, reducing monetisation to the freemium scheme alone risks being too narrow today. Increasingly, in fact, the economic model of AI platforms combines subscriptions, enterprise APIs, priority access to infrastructure and licensing agreements for data and content. In some sectors, especially audiovisual and music, one can even observe a transition from generalised scraping logics to training models based on proprietary or authorised catalogues, as shown by the Lionsgate-Runway partnership and the most recent licensing agreements in the music sector.[9899]

This transformation of AI into a cognitive infrastructure and economic platform does not produce effects only on the side of the production and circulation of content but is also beginning to affect the organisation of work and the structure of **professions**. Recent empirical evidence on the effects of artificial intelligence on the labour market seeks to assess the impact of AI models on work, distinguishing between their theoretical capability — that is, the set of tasks they could perform or accelerate in the abstract — and their actual use, namely the tasks in which such systems are concretely employed in professional contexts and with what intensity.[100]

[98]See: "'Hunger Games' studio Lionsgate announce AI video deal", from BBC.

[99]See: "Major labels' licensing deals with AI companies: ECSA calls for transparent licensing agreements that truly value the works of composers and songwriters", from the ECSA website.

[100]See, among others, Massenkoff, M., & McCrory, P. (2026). Labor market impacts of AI: A new measure and early evidence. Anthropic, and AI's Labor Impact, HAI The 2026 AI Index Report (2026). Stanford University.

Among the most significant findings is that the workers most exposed tend, more frequently, to be women, individuals with higher levels of education, older workers and better-paid workers. By contrast, occupations with a predominantly manual content appear, at least for now, to be less exposed, until the evolution of robotics makes a more direct impact on those activities possible as well (see Figure 6).

It follows that, at least in this initial phase, the risk is not concentrated in lower-skilled or lower-income jobs, but rather in a significant portion of intellectual and administrative work. This result, which may appear paradoxical, in fact derives from the specific characteristics of the "artificial intelligence product", which, as noted, is a powerful super-cognitive system and tends, in this sense, to replace, at least in part, the intellectual activity of workers (and not only workers).

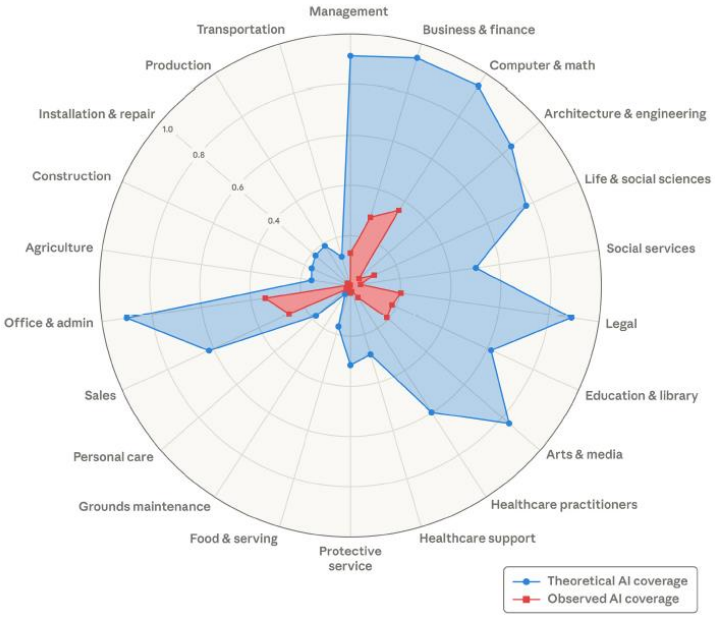


Figure 6 – Theoretical AI capability and observed AI exposure by occupational category[101]

Source: Labor market impacts of AI: A new measure and early evidence.

[101]The blue area represents the share of activities that large language models could theoretically perform; the red area indicates the coverage actually observed in Claude's professional-use data, aggregated by broad occupational categories.

Another relevant and not immediately obvious finding is the empirical observation that, at present, there is no wave of layoffs directly attributable to artificial intelligence. However, problematic signs are emerging on the side of new hiring: there is a slowdown in entry into the most exposed occupations, especially for workers aged between 22 and 25, particularly in junior profiles such as programmers, customer-service workers and data-entry operators. In general, workers in this age group employed in the occupations most exposed to AI show employment levels around 16% lower than workers in the same age group employed in less exposed occupations. The data therefore suggest that there is no clear evidence of an increase in unemployment resulting from a generalised and uniform substitution of employment, but rather a possible contraction in access opportunities to certain professions, especially for younger workers and in specific corporate functions.[102]

This breakdown between the potential capability of models and their actual adoption in production processes helps explain why the observable effects on aggregate employment are, at least for now, slower than the speed at which the technical capabilities of systems are evolving. In many cases, artificial intelligence does not fully replace an occupation, but affects specific portions of work, modifying its content, modes of performance and internal distribution of tasks. The result is a gradual but widespread restructuring process, with differentiated trajectories across cognitively intensive sectors, administrative activities, professional services and creative industries.

Read in this light, artificial intelligence cannot be treated simply as a tool: it is a new intermediary of knowledge that reorganises complex economic relationships among the production, appropriation and dissemination of information, all mediated by pricing mechanisms or by their absence.

4.3 Production structure

The production of artificial-intelligence systems — in particular large-scale generative systems — is based on a cost structure heavily skewed towards fixed and **sunk costs**, which make this technology profoundly **capital-intensive**. This characteristic is not neutral from a competition standpoint, but both reflects and reinforces a highly concentrated configuration of power, in

[102]See Figure 4.4.30 on page 222 of Stanford University's AI Index Report 2026.

which a small number of global operators control the critical infrastructures of compute, data and market access. The main cost items include:

- data collection, cleaning and labelling;
- the computationally intensive training of large-scale models;
- hardware infrastructure (GPUs and specialised accelerators such as TPUs, as well as high-density data centres);[103]
- research and development of new architectures and software optimisations.

In this context, the true centre of gravity of costs — and of the industrial power that derives from them — lies in compute, that is, the availability of advanced computing power required for the training and operation of models at scale. The centrality of this factor is clearly shown by the rapid expansion of the AI infrastructure base: since 2022, the global computational capacity attributable to the main AI chips has grown by around 3.3 times per year, reaching in 2025 the equivalent of 17.1 million H100s, i.e. high-performance chips taken as the benchmark for AI computing power[104]; in parallel, the power capacity of AI-dedicated data centres reached around 29.6 GW in the fourth quarter of 2025 (for an analysis of issues relating to the environmental impact of AI, see § 5.4). These data confirm that competition in AI is now played out through infrastructure investments on an exceptional scale, in relation to which compute today represents the main component of fixed costs and, at the same time, the main technological bottleneck of artificial intelligence. Its relative scarcity, the high capital intensity required and the time needed to build infrastructures make entry by new operators extremely difficult (for the environmental profiles associated with training and inference and with the growth of data centres, see § 5.4).[105]

Control of computing power unfolds across **two closely interdependent levels**:

- on the one hand, the design and production of **advanced chips**;
- on the other, the availability and management of the **physical infrastructures** in which those chips are installed and used, in particular hyperscale data centres.[106]

[103]Tensor Processing Unit (TPU): ASIC accelerators specialised for neural networks.

[104]Computing units comparable to the Nvidia H100 chip, now used as the reference benchmark for measuring the computational power employed in AI systems.

[105]The 2026 AI Index Report (2026).

[106]Cellini, P., Ibarra, M. (2024), AI Impact, Luiss University Press.

The **first level** — chip production — depends not only on industrial capacity, but is also conditioned by a material constraint: the availability of critical raw materials, including rare earths, necessary for the construction of data-centre hardware (electronic components, magnets, cooling systems and network infrastructures), which have highly concentrated supply chains, making them vulnerable to commercial or geopolitical disruptions: China, for example, controls around 98–99% of the supply of refined gallium.

Data-centre block	Where they enter the data centre	Critical raw materials / associated materials	Why it matters (AI-related driver)
Server boards and circuits	PCBs, connectors, soldering, internal cabling	Silver, gold, copper, tin, tantalum, palladium, nickel	More servers and higher electronic density → increases the amount of components and interconnections.
Semiconductors and microchips	CPUs/GPUs/accelerators, logic devices, microelectronic components	Silicon; gallium, germanium, indium and arsenic for compound semiconductors and photonics; fluorine/fluorinated compounds mainly as process materials	Growth in compute → greater demand for advanced chips and for base and process materials for microelectronics.
Thermal dissipation and structure	Heat sinks, exchangers, chassis, support parts	Copper, aluminium	More power = more heat to dissipate → increases the need for materials for cooling and mechanical structures.
Magnets and data storage	Motors/actuators (fans, pumps), HDD components and related parts	Rare earths (neodymium, praseodymium, dysprosium, terbium), boron, copper, aluminium	Need to manage thermal stress and reliability → efficient magnets for moving parts; "mass" storage on HDDs remains a significant materials driver.
Network and connectivity	Optical fibre, communication cables, intra/inter-data-centre interconnections	Germanium and, in some routes and optical equipment, erbium; as well as copper and aluminium for cabling and power supply	AI increases traffic and interconnection → high-speed connectivity becomes a structural constraint.

Table 5 – Critical raw materials for the construction of data centres

According to the IEA, the rapid growth of AI and data centres requires significant quantities of minerals and metals — including copper, aluminium, silicon, gallium and rare earths — and capacity expansion to 2030 may have a measurable impact on overall demand, with particularly relevant pressures on certain materials. In this context, the risk of bottlenecks along the materials supply chain becomes an integral part of the production structure of AI. Table 5 summarises how these raw materials enter the main functional blocks of a data centre and why they are strategic in light of the spread of AI.[107]

[107]International Energy Agency (2025), Energy and AI.

The **second level** concerns the ownership and management of "hyperscale" data centres: it is here that the market position of certain hardware and cloud-infrastructure providers translates into indirect control over the entire AI value chain, from model training to their distribution on the market. Consider that around half of the world's data centres are currently managed by cloud infrastructures attributable to a few major operators.[108] The geographical distribution of these infrastructures also shows a high degree of concentration: according to the 2026 AI Index, in 2025 the United States accounted for 57.1% of the data centres considered in the figure, compared with 5.6% for Germany and 5.5% for the United Kingdom; for Italy, the share stood at around 1.8%.[109] The cloud-services market is dominated by Amazon (AWS), Microsoft (Azure) and Google Cloud, which together control more than 60% of the global market.[110] These operators do not merely provide computing capacity to third parties, but constitute the main gateway to the computational resources needed to develop advanced models, to which formally distinct entities such as OpenAI or Anthropic are also tied, since their research and development capabilities structurally depend on the infrastructures made available by the major technology operators[111] (see also § 4.4). If Chinese control over rare earths affects the material availability of hardware upstream, the control exercised by the United States over the advanced-semiconductor supply chain operates at a technological and systemic level, affecting the very conditions of access to compute. This position is based on primacy in chip design and on the centrality of US companies in the highest-value-added segments. Downstream of compute, this structure is reflected in a high degree of industrial concentration: NVIDIA holds an estimated share of around 80% of chip design for data centres[112], while the production chain for advanced semiconductors is itself extremely concentrated: TSMC produces around 90% of the world's

[108]In this regard, see: Cellini (2024), p. 96.

[109]See Stanford University's AI Index Report 2026, which reports the following absolute values in terms of data-centre distribution: United States 5,427; Germany 529; United Kingdom 523; China 449; Canada 337; France 322; Australia 314; Netherlands 298; Russia 251; Japan 222; Brazil 197; Mexico 173; Italy 168; India 153; Poland 144.

[110]Specifically, AWS's share ranges between 40-62%, Azure's between 10-35% and Google Cloud's between 5-10%. In this regard, see: Organisation for Economic Co-operation and Development - OECD (2025), Competition in artificial intelligence infrastructure, OECD Roundtables on Competition Policy Papers, No. 330, OECD Publishing, Paris.

[111]Aresu, A. (2024). Geopolitics of Artificial Intelligence. Feltrinelli Editore.

[112]Organisation for Economic Co-operation and Development - OECD (2025), Competition in artificial intelligence infrastructure, OECD Roundtables on Competition Policy Papers, No. 330, OECD Publishing, Paris, https://doi.org/10.1787/623d1874-en. Specifically, see the following passage: "Nvidia (a fabless company) has emerged as the market leader in the sector, with recent estimates suggesting that the firm has over 80% market share for GPU chips used for AI".

most advanced chips[113] and ASML is the only global supplier of the EUV machines needed for their fabrication. Figure 7 makes immediately visible the concentration (including geographical concentration) of the semiconductor and memory supply chain, helping to explain the vertical interdependence and strongly asymmetric cost structure discussed below. 4.4[114]

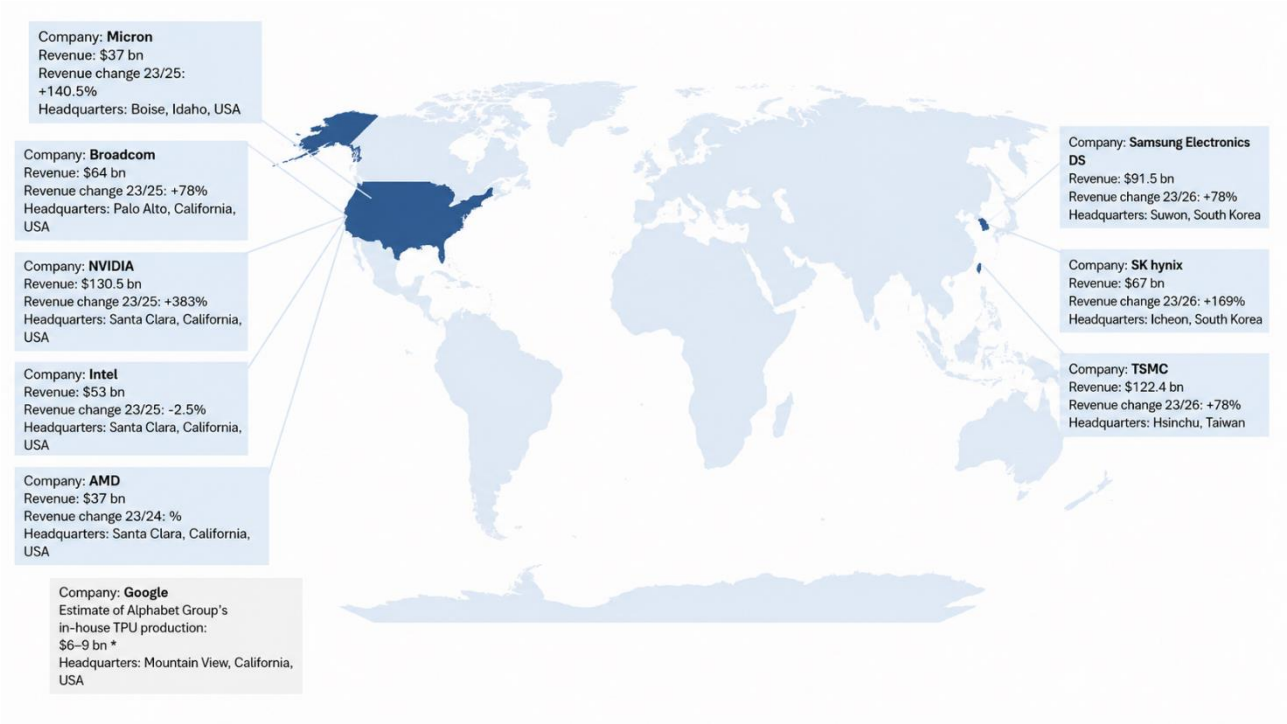


Figure 7 – Main semiconductor companies for AI5

** For Google, the figure indicated does not represent revenues, but an external estimate of TPU spending in 2024, not directly obtainable from Alphabet's financial statements. See in this regard: "Google preparing to partner with Taiwan's MediaTek on next AI chip, Information reports", from Reuters.*https://www.reuters.com/technology/artificial-intelligence/google-preparing-partner-with-taiwans-mediatek-next-ai-chip-information-reports-2025-03-17/

The combination of concentration in chip design, chip production and data-centre management increases the market power of vertically integrated operators, with a strongly asymmetric cost structure between those actors and other companies, thereby reducing the contestability of the sector. Control of the critical resources of compute, together with the position held at the various stages of the supply chain, allows integrated operators to sustain — and partly

[113]Organisation for Economic Co-operation and Development - OECD (2025). Specifically, see the following passage: "No other firm has been able to commercialise an alternative to ASML's Extreme Ultraviolet (EUV) Lithography technology that is necessary to manufacture the latest generations of chips."

[114]Cellini (2024), p. 95.

internalise — extremely high fixed costs, which are instead prohibitive for potential new entrants.

Increasingly often, what makes this exceptional level of investment economically sustainable is a specific financial architecture based on circular agreements among the main actors in the artificial-intelligence ecosystem.

Box 4 – The power of compute: CPU, GPU, TPU

In recent years, GPUs (Graphics Processing Units) have moved from being components for graphics and video games to central computational infrastructures for artificial intelligence, particularly for deep learning, which requires the execution of an enormous number of mathematical operations in parallel on large quantities of data. It is in this context that GPUs have proved decisive.

CPUs (Central Processing Units) remain fundamental for general and coordination tasks, but they are not optimised for the massively parallel computation required for model training. GPUs, by contrast, sacrifice a degree of flexibility to achieve extreme parallelism, thanks to thousands of simple cores operating simultaneously, making them particularly suited to the training and execution of neural networks.[115][116]

On this basis one can explain the rise of NVIDIA, which today occupies an almost hegemonic position in the open market for AI computing infrastructures. Its advantage depends not only on hardware, but above all on the software ecosystem (CUDA), which has created a strong technological dependence. NVIDIA is no longer merely a chip producer, but a provider of cognitive infrastructures, making the availability of GPUs a global strategic factor.

Google's strategy is different: it has developed proprietary chips, the so-called TPUs (Tensor Processing Units), highly specialised for machine learning. Less versatile than GPUs but more efficient for specific workloads, TPUs strengthen a model of vertical integration, in which control of the hardware supports the company's entire ecosystem of AI services and models.[117]

Alongside these two poles, AMD is NVIDIA's main competitor on the hardware side, but lags behind in the software ecosystem, while Intel, although central in data centres, is not the reference point for the training of the most advanced models.

[115]Federal Trade Commission, FTC (2025). Partnerships between cloud service providers and AI developers. FTC staff report on AI partnerships & investments.
[116]See: "China doubles down on microchips: a new 47.5-billion-dollar fund established", from Forbes.
[117] The TPU is an AI accelerator designed by Google specifically for machine/deep learning. Unlike the GPU, the TPU is an ASIC (*Application-Specific Integrated Circuit*), i.e. a chip built to do one specific task: handling "tensors", the data structures underlying learning. TPUs optimise data flows for linear-algebra operations, reducing training times and energy consumption. However, they are less flexible, because they excel almost exclusively in areas linked to specific frameworks and large data centres.

Apple, finally, follows a distinct trajectory, focusing on accelerators integrated into its own chips to develop AI models directly on the device, rather than on industrial AI.

Overall, a new geography of computational power is emerging: GPUs and specialised chips are no longer simple technical components, but strategic infrastructures with economic, geopolitical and regulatory implications. Understanding the differences between CPUs, GPUs and TPUs today means understanding who controls the capacity to produce artificial intelligence and under what conditions.

In this configuration, high infrastructure costs are not borne by individual operators in isolation, but are distributed along a network of cross-relationships among model developers, chip suppliers and cloud operators, through mechanisms that combine equity investments, multi-year supply contracts and advance spending commitments for compute services. These "circular agreements" make it possible to transform a significant share of fixed costs into financial flows internal to the system, reducing individual risk and, at the same time, reinforcing interdependencies among the actors involved.

From an economic perspective, these agreements perform a crucial function in supporting demand for computing power, enabling model developers to access computational capacity that would otherwise be unsustainable and allowing infrastructure and hardware providers to secure sales volumes sufficient to justify investments on an exceptional scale. The result is a self-reinforcing circuit in which capital, infrastructure and revenues mutually strengthen one another, contributing to the rapid expansion of the ecosystem while at the same time accentuating barriers to entry and systemic dependence among a small number of major operators.

In this context, some agreements among model developers, cloud-infrastructure providers and chip producers — summarised by way of example in Table 6 — show how financial commitments are steadily in the order of tens or hundreds of billions of dollars over multi-year horizons, confirming the highly capital-intensive nature of the sector.*Table 6 – Main agreements among AI operators: nature of commitments and network of interdependencies5*

Useful empirical evidence for characterising this contractual architecture is provided by a study on partnerships between large cloud providers and generative-model developers recently carried out by the US Federal Trade Commission. The report highlights that, in the main partnerships examined, clauses recur — although not to the same extent — combining: (i)

equity stakes and/or revenue-sharing mechanisms benefiting the cloud partner; (ii) rights of consultation, influence and, in some cases, control, as well as exclusivity and preference provisions, including with respect to model-release timelines; (iii) cloud-spending commitments requiring the developer to allocate a significant share of the investment received to the services of the cloud partner (so-called circular spending); (iv) the sharing of resources and information, such as discounted access to computing capacity, chip co-development plans, exchanges of technical personnel and access to financial, performance and infrastructure-requirement data. The FTC also identifies, as "areas to monitor", the possible effects of such clauses on access to essential inputs (compute and knowledge/talent), on the increase in switching costs, both contractual and technical, and on the informational advantages deriving from the flow of sensitive data between partners. Overall, the report supports the idea that these partnerships are not limited to a mere financial investment or an ordinary customer-supplier relationship, but may give rise to forms of contractual integration capable of strengthening lock-in, information asymmetries and interdependencies along the entire AI supply chain.[118]

Table 6 sheds light on the interdependencies that emerge in the form of commercial and corporate agreements among US private operators. It does not, however, include the main Chinese actors, whose development is more often also supported by public funds and industrial-policy instruments (as shown by the establishment of the 344-billion-yuan semiconductor fund), according to logics that are often not directly comparable with market-based ones.[119]

Parties involved	Description of the agreement	Declared / estimated value
OpenAI ↔ Microsoft	Microsoft has invested in OpenAI (equity + cloud credits); OpenAI uses Azure as the primary infrastructure for model training and deployment.	Microsoft owns 27% of OpenAI.[120] Multi-year cloud-spending commitments are estimated at between USD 200 and 250 billion (long-term contracts).[121]

[118] Federal Trade Commission, FTC. (2025). *Partnerships between cloud service providers and AI developers*. FTC staff report on AI partnerships & investments.

[119] See: "*La Cina punta ancora sui microchip: istituito un nuovo fondo da 47,5 miliardi di dollari*", Forbes.

[120] See: "*OpenAi diventa* società *a scopo di lucro. Microsoft rileva il 27% e supera i 4mila miliardi*", La Stampa.

[121] See: "*The next chapter of the Microsoft–OpenAI partnership*", da Microsoft Corporate Blogs.

Parties involved		Description of the agreement	Declared / estimated value
OpenAI Nvidia	↔	Nvidia supplies advanced GPUs; commercial relationships and technological cooperation. There is no public evidence of direct Nvidia investments in OpenAI, but there is a strong cross-commercial dependence.	USD 30 billion[122] (this amount has been revised. The previous agreement provided for an investment of USD 100 billion).
OpenAI Oracle	↔	Oracle builds and manages dedicated data centres for AI workloads; OpenAI commits to using the capacity.	Multi-year contracts estimated at around USD 300 billion, with launch expected in 2027.[123]
OpenAI Amazon (AWS)	↔	Infrastructure-use agreements for specific workloads; a non-exclusive relationship complementary to Azure.	Amazon announced an investment in OpenAI of up to USD 50 billion + a 2 GW compute commitment.[124]
OpenAI Google	↔	Infrastructure and cloud agreement between Google and OpenAI, under which OpenAI added Google Cloud among its compute-capacity providers to support model training and inference, in a collaboration defined in 2025 despite direct competition between the two companies.[125]	At present, there is no evidence of a Google equity investment in OpenAI comparable to Google's investment in Anthropic.
OpenAI AMD	↔	OpenAI will purchase 6 gigawatts from AMD for new data centres. Agreement on OpenAI's right to purchase AMD shares.	Value not public; technological agreement (for 6 gigawatts): agreement on OpenAI's right to purchase 10% of AMD.[126]
Anthropic Amazon (AWS)	↔	AWS is Anthropic's main cloud and training partner; Anthropic also uses Trainium and Inferentia chips and distributes Claude models through Amazon Bedrock.	USD 4 billion total investment; further long-term cloud/compute commitments are not fully public.[127]
Anthropic Google	↔	Google provides cloud and infrastructure support to Anthropic, which uses Google Cloud and TPUs for training and services.	In 2023, USD 2 billion of investment by Google in Anthropic was made public; Google now holds about 14%.[128] At the end of 2025, Anthropic announced an expansion of its use of Google Cloud and up to 1 million Google TPUs; the transaction is described as worth tens of billions of dollars and with more than 1 gigawatt of expected capacity in 2026.[129] [130]
Anthropic Microsoft	↔	Infrastructure partnership on Azure, with Claude integrated into the Microsoft ecosystem.	Microsoft announced an investment in Anthropic of up to USD 5 billion.[131]

[122] See: *"OpenAI's $110 billion funding round draws investment from Amazon, Nvidia, SoftBank"*, Reuters.

[123] See: "*Accordo OpenAI-Oracle, 300 miliardi di investimenti in potenza di calcolo in 5 anni*", Sole24Ore.

[124]See: "*OpenAI's $110 billion funding round draws investment from Amazon, Nvidia, SoftBank*", da Reuters.

[125]See: "Exclusive: OpenAI taps Google in unprecedented cloud deal despite AI rivalry, sources say", Reuters..

[126] See: "*OpenAi sceglie Amd, mega ordine di chip e 10% delle azioni: come cambia la sfida sull'intelligenza artificiale*", dal Corriere della Sera.

[127]See: "Amazon concludes $4 billion investment in Anthropic", Amazon News.

[128]See "*Google agrees to invest up to $2 billion in OpenAI rival Anthropic*", Reuters.

[129]See: "*Google has given Anthropic more funding than previously known, show new filings", TechCrunch.*

[130]See "*Anthropic to use Google's AI chips worth tens of billions to train Claude chatbot*", Reuters.

[131] See: "*Microsoft, Nvidia to invest in Anthropic as Claude maker commits $30 billion to Azure*", con riferimento al passaggio: "Nvidia (NVDA.O), opens new tab will commit up to $10 billion to Anthropic and Microsoft (MSFT.O), opens new tab up to $5 billion", Reuters.

Parties involved	Description of the agreement	Declared / estimated value
Anthropic ↔ Nvidia	Strategic partnership for infrastructure and scaling of Claude models on Nvidia systems; an investment by Nvidia in Anthropic was also announced.	Nvidia announced an investment in Anthropic of up to USD 10 billion.[132]
Nvidia ↔ Amazon (AWS)	Industrial agreement on AI chips and networking: Nvidia will supply AWS with up to 1 million GPU chips by 2027, as well as networking components and other technologies for AI infrastructure.[133]	The financial terms of the agreement have not been made public.
Nvidia ↔ xAI	GPU supply agreements (equity investment conditional on hardware purchases).	Possible equity investment conditional on hardware purchases. The estimated value of the transaction would be around USD 2 billion (estimate).[134]
Nvidia ↔ CoreWeave	Nvidia is a significant shareholder and the main GPU supplier; CoreWeave uses Nvidia hardware almost exclusively.	USD 6.3 billion in compute supply + USD 2 billion investment in CoreWeave shares by Nvidia.[135]
OpenAI ↔ CoreWeave	GPU computing-power supply contracts; CoreWeave granted equity to OpenAI.	Total contracts amounting to USD 22.6 billion.[136]
CoreWeave ↔ Meta	Infrastructure agreement for AI capacity: CoreWeave signed a contract with Meta to provide AI computing/cloud capacity until 2032.[137]	The value of the agreements is USD 14.2 billion (September 2025 agreement) + USD 21 billion (April 2026 expansion agreement), for a cumulative total of around USD 35.2 billion.

Table 6 – Main agreements among AI operators: nature of commitments and network of interdependencies5

This financial architecture therefore represents the financial complement to the technological concentration described above: if chip design, chip production and the management of hyperscale data centres constitute the industrial backbone of compute, circular agreements constitute its implicit financing mechanism, making it possible to sustain an infrastructure-investment dynamic unprecedented in volume, intensity and speed.

The cost levels observed in the training phase of large models should be placed within this framework. By way of example, training the GPT-3 model (175 billion parameters) involved an estimated cost of about USD 4.6 million for computation alone, while for more recent models —

[132]See: "*Microsoft, Nvidia to invest in Anthropic as Claude maker commits $30 billion to Azure*", con riferimento al passaggio: "Nvidia (NVDA.O), opens new tab will commit up to $10 billion to Anthropic and Microsoft (MSFT.O), opens new tab up to $5 billion", Reuters.

[133]See:"*Nvidia to sell 1 million chips to Amazon by end of 2027 in cloud deal*", Reuters.

[134]See: "*Musk's xAI nears $20 billion capital raise tied to Nvidia chips, Bloomberg News reports*", Reuters.

[135]See: China's AI Companies Are Going Closed Source.

[136]See: "*CoreWeave inks $6.5 billion deal with OpenAI*", da CNBC.

[137]See: "*CoreWeave signs $14 billion AI infrastructure deal with Meta*" e "*Meta, CoreWeave deepen AI cloud partnership with fresh $21 billion deal*", da Reuters.

such as GPT-4 — estimates exceed USD 60 million and reach over USD 100 million for some new-generation frontier models.[138] The energy consumption associated with these operations is also very high (see § 5.4). [139] According to estimates developed over the years, the training costs of generative-AI models have shown a steadily rising trend. In 2017, training the original transformer model — the basis of most current LLMs (see § 3.2.2) — cost about USD 670. In 2019, RoBERTa Large required an investment of around USD 160,000, while in 2023 the estimated cost of training OpenAI's GPT-4 reached about USD 79 million. The estimated costs for Llama in 2024 are even higher, at around USD 170 million (Figure 8). Today, the most advanced models tend to disclose ever less information on parameters, datasets and training duration, making it more difficult to estimate their actual costs precisely and, more generally, to assess transparently the conditions under which the most powerful systems are developed.[140]

If all costs associated with the creation and development of AI models are considered (training hardware, energy, cloud, R&D, staff), it has been calculated that the total cost of the latest-generation models will exceed one billion dollars by next year.[141]

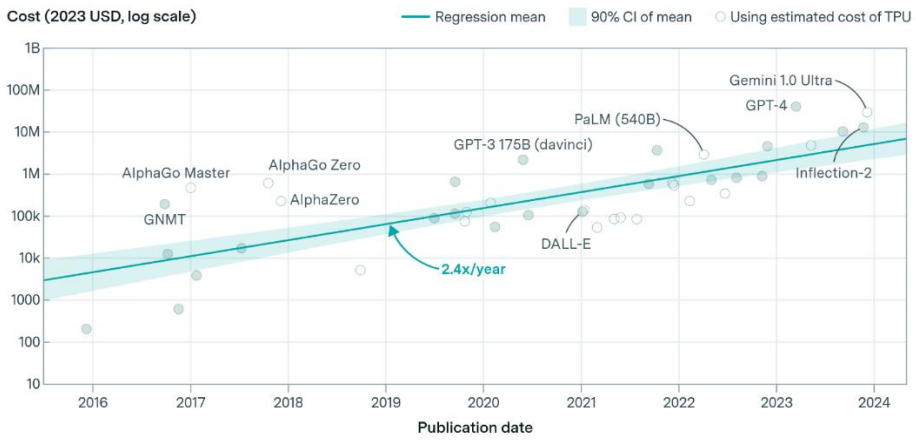


Figure 8 – Training cost (hardware and energy) of generative-AI models

Source: Epoch AI

[138]Cottier, B. (2023), Trends in the Dollar Training Cost of Machine Learning Systems.

[139]Training GPT-3 required the equivalent of the electricity consumed annually by more than 1,000 households, and a single day of ChatGPT use may be equivalent to the daily needs of 33,000 US households.

[140]The 2025 AI Index Report (2025). Stanford University, Human Centered Artificial Intelligence – HAI.

[141] Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. *arXiv preprint arXiv:2405.21015*.

These fixed costs generate significant returns to scale (and barriers to entry): once a model has been developed, the marginal cost of serving it to a growing number of users is relatively low, which encourages concentration of supply.

Moreover, there is an empirically observable correlation between the amount of data used, computing power, algorithmic progress and the predictive performance of models, which reinforces the cumulative advantages of incumbent operators (Figure 9).[142]

These dynamics are complemented by the contribution of end users, who, as illustrated in § 3.1, indirectly contribute to the improvement of models through iterative learning mechanisms (feedback, corrections, usage data).

Taken together, these factors give rise to **returns to scale** both on the supply side (technological scale and algorithmic learning: economies of scale) and on the demand side (more users, greater value for each user: network effects).

Figure 9 – Estimates of the contributions of compute scaling and algorithmic innovation to achieving state-of-the-art performance (the contribution of algorithmic progress is about half that of compute scaling)

Source: Epoch AI

[142] Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., ... & Sevilla, J. (2024). Algorithmic progress in language models. *Advances in Neural Information Processing Systems*, *37*, 58245-58283.

However, in recent years these barriers have been partially attenuated by the evolution of new techniques. The trend towards more compact and efficient models is now consolidated: DeepMind has shown that training smaller models on larger datasets can produce results at least comparable to larger models trained on less curated data, confirming that algorithmic efficiency and data quality can, at least in part, substitute for mere computational scale. At the same time, technical and operational solutions have spread that make it possible to compress average costs and reduce the minimum efficient scale required to compete, thereby broadening — at least downstream — the spaces for market entry.

In this framework, the infrastructural dynamics of compute are overlaid by competition among "technological blocs", understood as industrial ecosystems and de facto standards (models, cloud platforms, software stacks, distribution channels). In an initial phase, large language models were developed mainly according to a proprietary and "closed" approach, in which capabilities were protected as trade secrets and made available primarily via cloud: this allowed — and partly still allows — pioneer operators to monetise their time advantage, as well as to retain demand by keeping users within their own proprietary ecosystem (the so-called lock-in effect).

In a later phase, a strategy emerged, especially among some operators with still limited market shares, based on open-weight models and forms of controlled openness, aimed at reducing the advantage accumulated by closed systems through faster and more widespread diffusion and greater accessibility of the technology, while not removing the upstream structural constraint represented by compute. Open-weight models use openness to accelerate standardisation, expand the developer base and foster new opportunities for integration with services used by end users. This is a recurring dynamic in digital markets: the "open" (or semi-open) standard can serve as a competitive lever to weaken pre-existing closed ecosystems (as in Android vs iOS, Linux vs Windows or ROCm vs CUDA), with the effect of reshaping competitive balances between those who control the platform and those who control the diffusion and large-scale use of the technology.

In this scenario, Stanford University's AI Index Report 2026 shows that competition among models is no longer played out only on performance, but increasingly also on access modalities and degrees of transparency. In other words, competitive advantage depends not only on how

efficient a model is, but also on who controls access to it, the delivery infrastructure and the information needed to reproduce it.

It should be clarified that "accessibility" of models does not mean the simple possibility for ordinary users to use a generative-AI system through an interface, but rather the possibility for developers and professional operators to access the model as a technological infrastructure to be integrated into their own services, applications and processes. From this perspective, API access, the release of weights or the non-publication of training code describe different degrees of technical availability of the model and different levels of dependence on the provider.

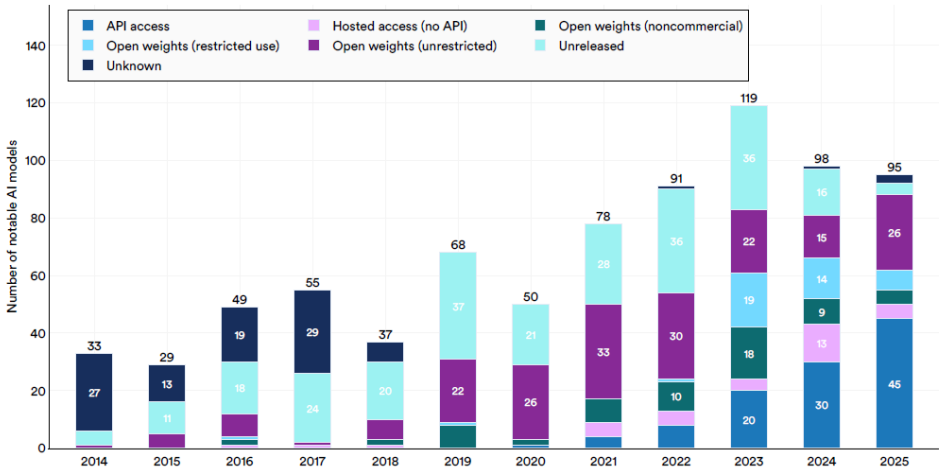


Figure 10 – Evolution of release modalities for notable AI models (2014–2025)

Source: AI Index Report 2026

The fact that in 2025 API access was the most frequent release modality (45 models out of 95; see Figure 10) indicates that many models are made available not as fully transferable technical goods, but as provider-intermediated services: developers, firms and other professional users can use them for their own applications and services, but under conditions defined unilaterally by the platform operator, which controls prices, usage limits, filters, logging and cloud integration.

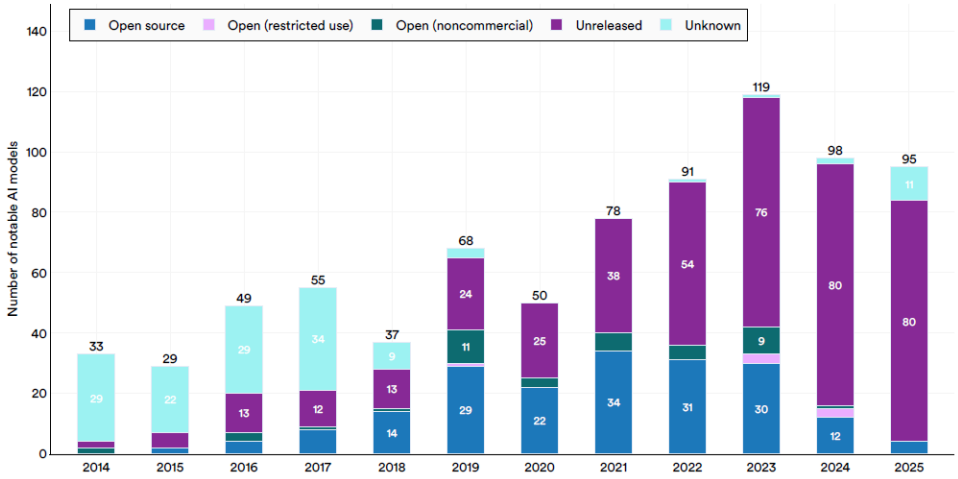


Figure 11 – Evolution of training-code accessibility in AI models

Source: AI Index Report 2026

Added to this is a second datum, even more relevant from a structural standpoint: 80 models out of 95 did not make their training code available (Figure 11). This means that, even when a model is accessible or partially open, the possibility of fully reconstructing its development process, verifying its training methods, replicating its results or subjecting it to independent audit often remains precluded. It follows that the "openness" of open-weight models is indeed an important competitive lever, but it operates within a general context in which frontier systems tend to be increasingly opaque and increasingly controlled by the major operators that jointly control model, cloud, distribution and user interface.

Box 5 – Average AI costs and new techniques

In recent years, operators have used a number of techniques aimed at improving model performance and reducing average training costs for algorithms and service provision. The main ones are:

- **Model distillation:** A technique that transfers knowledge from a large model (teacher) to a smaller one (student), drastically reducing computational load

while maintaining comparable performance, with a significant reduction in inference as well as training costs.
Example: DeepSeek-V3 offered performance close to GPT-3.5 with a fraction of the required power.

- **Fine-tuning:** Allows smaller operators to specialise pre-trained models on specific domains (healthcare, legal, customer service), reducing the costs of the entire training cycle.
  Example: fine-tuning Mistral for legal uses may cost less than USD 100,000, compared with several million for training from scratch.
- **Open-source and open-weight models:** Projects such as LLaMA, Mistral, DeepSeek, or developments by EleutherAI and Hugging Face have opened access to the technology, reducing dependence on incumbents. This also enables public laboratories or private start-ups to develop AI services with limited resources, favouring greater downstream contestability, despite persistent upstream concentration.

These technologies significantly lower total average costs and the minimum scale required to operate, opening spaces for competition and innovation even for actors with smaller budgets. However, the competitive advantage in terms of infrastructure, proprietary data and distribution remains largely concentrated in the hands of a few global operators.

It is no coincidence that many operators — first OpenAI/Microsoft and subsequently Anthropic/Microsoft, Google/DeepMind, Amazon, Alibaba and Meta — operate under vertical integration, simultaneously controlling many stages of the artificial-intelligence supply chain: data and information, models, hardware infrastructures, market access and user interfaces.

In summary, the production structure of AI reflects a capital-intensive economy, fuelled by internal feedback and external networks, in which technological efficiency and economic sustainability depend on the ability to reduce average costs, above all by exploiting scale advantages.

4.4 Services, operators and markets

We have seen that the AI market is characterised by high technical barriers (algorithmic innovations), economic barriers (high fixed and sunk costs) and strategic barriers (vertical and diagonal integration). In particular, the segment of large-scale generative models is a **concentrated and strategically integrated sector**. A handful of major operators — predominantly US-based — dominate the development, distribution and integration of AI into

digital systems, benefiting from economies of scale, increasing returns and cumulative advantages that are difficult to replicate. This leads to a strongly oligopolistic industrial structure, in which technological innovation is closely intertwined with control over global digital platforms.[143]

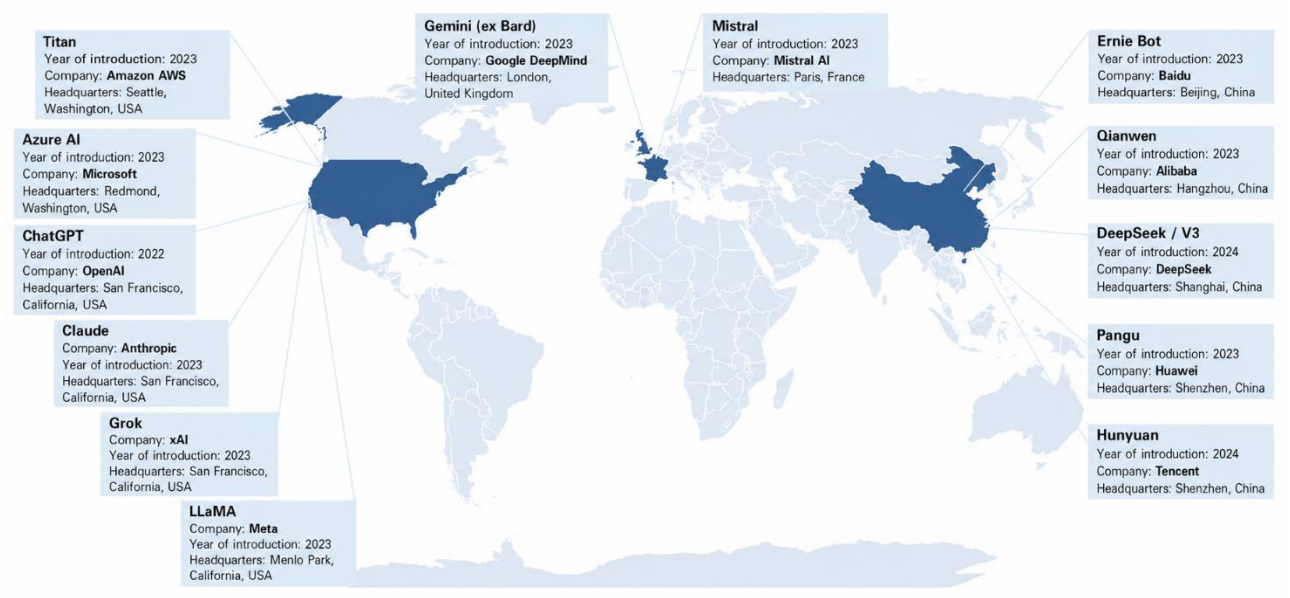


Figure 12 – Generative-AI services worldwide

The main actors in the sector — OpenAI, Google DeepMind, Amazon AWS, Meta, xAI — are not merely producers of technologies, but architects of entire AI-based digital ecosystems. Each has developed a proprietary model, a set of models, or entered into commercial or financial agreements (to explore the network of interdependencies among operators, see Table 6):[144]

- OpenAI is deeply intertwined with Microsoft (the latter has invested tens of billions of dollars in OpenAI, also providing its Azure cloud service, and in return holds a minority stake in the company; however, the relationship is no longer characterised by the

[143] Recent market research not only highlights a surge in investment in generative AI — suffice it to note that, in 2024, this segment accounted for around one quarter of the global AI market, exceeding USD 184 billion, with a compound annual growth rate of 24.4% projected between 2023 and 2030 — but also points to a significant shift in consumer habits: in 2023, as many as 13 million U.S. adults chose generative artificial intelligence as their primary tool for conducting online searches, a figure expected to exceed 90 million by 2027.

[144] OpenAI was initially linked to Microsoft through an exclusive cloud partnership, which was later progressively softened by the evolution of the agreement toward more multi-cloud arrangements and by Microsoft's integration of Anthropic models into Microsoft 365 Copilot as well.

original infrastructure exclusivity, and Microsoft has also integrated Anthropic models into Microsoft 365 Copilot);

- Google DeepMind, with Gemini, focuses on integration with Workspace and Android; moreover, Google/Alphabet has invested several billion dollars in Anthropic, also providing cloud services and holding a stake of about 14% in the AI company;
- Amazon AWS offers the Titan model and, more recently, also the Nova family, within its cloud offering for enterprises through Bedrock; at the same time, it has strengthened its alliance with Anthropic, becoming its main cloud and training partner;
- Meta offers the LLaMA family, which follows an open-weight strategy, favouring adoption by the open-source community; more recently, however, it launched Muse Spark, the first model in the new Muse series and the first model from the Meta Superintelligence Labs team, intended to power Meta AI and be progressively integrated into the services and products offered by the parent company;
- xAI, founded by Elon Musk, integrates the Grok model into the X platform (formerly Twitter);
- Mistral, the only significant European exception, offers highly efficient and fully accessible models.

In parallel, China has built its own national ecosystem, fuelled by giants such as Alibaba (Qianwen), Baidu (Ernie Bot), Tencent (Hunyuan) and Huawei (Pangu), alongside models such as DeepSeek, which stand out for their computational efficiency and open-weight approach. These actors are part of an industrial system oriented towards technological self-sufficiency and supported by active public policies.

In this context, Chinese models have also significantly increased their competitive weight: although China continues to rank behind the United States in the number of notable AI models released in 2025 (30 versus 50), the performance gap between the best models has now narrowed to very small values (Figure 13).

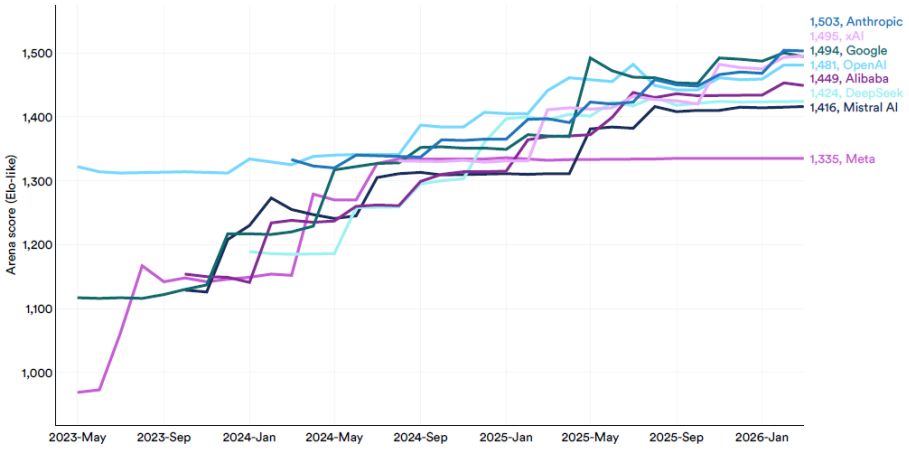


Figure 13 – AI performance differentials: United States vs China

Source: AI Index Report 2026

Already in February 2025, DeepSeek-R1 had temporarily equalled, and for a short time surpassed, the best US model in the Arena ranking, which measures model performance; in March 2026, the highest-ranked US model maintained an advantage of only 39 Arena points over the best Chinese model, corresponding to a 2.7% difference in score. The data therefore indicate not the elimination of every technological difference between the two countries, but the emergence of much closer competition at the top. The narrowing of the gap between the leading US and Chinese AI models signals the emergence of increasingly intense strategic competition, given the importance of the sector for a nation's entire economic, industrial and social system.[145-146-147]

One of the most relevant features of the large-scale artificial-intelligence sector is the degree of integration achieved by the main operators (see Figure 14).

The same companies that develop advanced AI models are often also:

i) Managers of market-access platforms (see § 4.2)

The same companies that develop AI also control the platforms that intermediate end-user access to the models. For example:

[145] See: Arena Leaderboard Dataset.

[146] See Figure 2.1.3 on page 77 of Stanford University's *AI Index Report*.

[147] The *AI Index Report 2026* states that, in February 2025, DeepSeek-R1 had temporarily matched, and briefly surpassed, the leading U.S. model in the Arena ranking; in March 2026, the highest-ranked U.S. model maintained a lead of only 39 Arena points, equal to 2.7%, over the leading Chinese model.

- Google distributes the Gemini model through its search engine, Android assistant and Google Workspace suite (Gmail, Docs, Meet). AI is also integrated into mobile devices (Pixel) and accessible via Bard/Gemini AI.
- Microsoft integrates GPT (from the OpenAI family) and Anthropic models into the Microsoft 365 suite through Copilot (Word, Excel, Teams) and into Azure AI cloud services. It thus has a vertical chain: models, interface and infrastructure.
- Amazon offers its Titan models through AWS Bedrock, targeting enterprises and developers. AI is also integrated into Alexa and AWS enterprise systems.
- xAI (Elon Musk's company) has integrated the Grok model directly into the X platform (formerly Twitter), positioning AI as a personalised assistant within the social network.
- Meta, initially focused on the technical distribution of LLaMA (through Hugging Face, AWS, Azure), has now launched its own conversational AI assistant, Meta AI, available directly in WhatsApp, Messenger and Instagram (via chat).

  Based on optimised versions of LLaMA 2 and LLaMA 3, this assistant allows users to ask questions, generate texts, obtain suggestions and create images ("Imagine with Meta AI"), leveraging Meta's vast user bases.

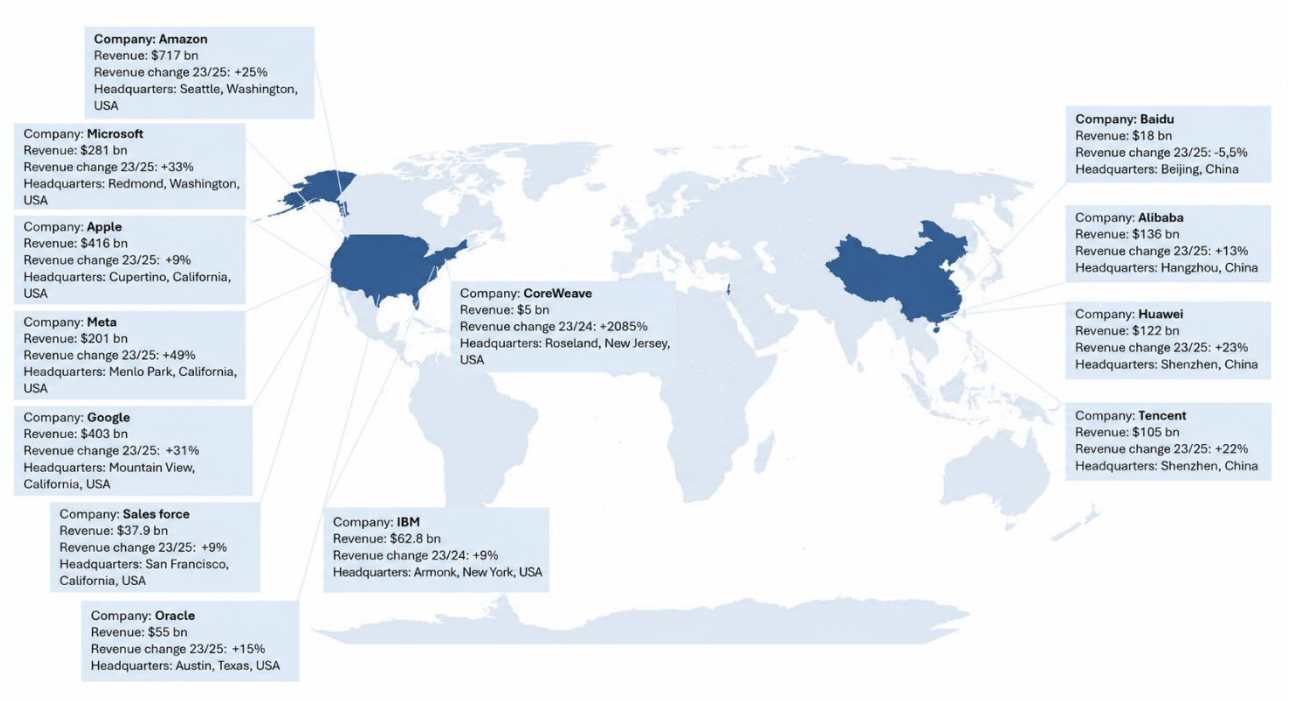


Figure 14 – Main operators in the AI field

ii) Computational-infrastructure providers (vertical integration)

Big Tech companies often have proprietary data centres and direct access to essential hardware components (see also Box 4). For example:Box 4

- Microsoft manages the Azure infrastructure, which hosts OpenAI and Anthropic models.[148]
- Amazon owns AWS, the world's leading cloud-services provider, on which the Titan model runs.
- Google uses its own network of TPUs (Tensor Processing Units), designed internally for training models such as Gemini.
- Meta has built data centres optimised to support the training and deployment of LLaMA models.

iii) Distributors of final digital services (diagonal integration)

AI is embedded in value-added applications and services, which feed direct and indirect revenue streams. For example:

- Meta, Amazon, Google, X and Microsoft have all adopted a strategy of integrating their AI into the package of services offered (with particular reference to Google);
- OpenAI offers ChatGPT as a subscription service, integrated into Copilot and other interfaces;
- Anthropic distributes Claude through APIs and partnerships with third-party developers and companies, including Google in particular.

This multi-level integration provides significant advantages in terms of efficiency, cost control, data access and speed of diffusion, but also increases the ecosystem's dependence on a few dominant actors. Barriers to access are not limited to model technology, but include the entire distribution, monetisation and feedback system that accompanies the use of artificial intelligence at scale.

4.5 Concluding remarks

Artificial intelligence is now widely recognised as a general-purpose technology (GPT), namely a technology with pervasive effects on production, economic and social organisation, and the

[148] See: "Microsoft, Nvidia Pump Billions Into Anthropic", Bloomberg.

generation of downstream innovations in a plurality of sectors. Like electricity, the computer or the Internet in the past, AI has the capacity not only to increase productivity in existing sectors, but also to radically transform the ways in which people work, consume and make decisions, in both the private and public spheres.

General purpose technologies, as Helpman and Trajtenberg explain, are "engines of growth" not so much because of their intrinsic performance, but because of their potential **complementarity** with other factors (organisation, human capital, regulation) and their ability to generate successive waves of innovation. AI fits fully into this logic: its applications — from natural-language generation to predictive analysis, from robotic control to logistics optimisation — extend horizontally across all sectors and vertically through all levels of the value chain.[149]

According to the McKinsey Global Institute, the widespread adoption of generative artificial intelligence could generate an annual economic impact of between USD 2.6 and 4.4 trillion globally, a figure comparable to the entire GDP of the United Kingdom or Germany.[150] The Organisation for Economic Co-operation and Development (OECD) has found that more than 70% of firms in advanced countries consider AI a priority technology for the next five years.[151] The World Economic Forum has highlighted AI's transformative potential in healthcare, energy, education and public administration, while also stressing the need for multilevel governance to manage systemic risks.[152]

At the social level, AI appears as a super-cognitive infrastructure: it mediates the way information is produced, aggregated and assessed; it shapes the perception of reality; and it automates decision-making processes at different levels. This opens significant opportunities in terms of access to knowledge and personalisation of services, but also entails new

[149] Helpman, E., & Trajtenberg, M. (1998). *A time to sow and a time to reap: Growth based on general purpose technologies*. In E. Helpman (Ed.), *General Purpose Technologies and Economic Growth*. MIT Press.

[150] McKinsey Global Institute (2023). *The economic potential of generative AI: The next productivity frontier.*

[151] Organisation for Economic Co-operation and Development – OECD (2023). *Artificial Intelligence Outlook 2023: Enabling Trust and Innovation*; Organisation for Economic Co-operation and Development – OECD (2024), *Digital Economy Outlook 2024 (Volume 1): Embracing the Technology Frontier*.

[152] World Economic Forum (2025), *AI in Action: Beyond Experimentation to Transform Industry*.

distributive risks relating to skills polarisation, professional disintermediation and concentration of informational power.

In this sense, artificial intelligence is not simply a new technological wave: it is a cognitive and organisational multiplier with systemic implications for the economy, society and democracy.

However, while AI's potential is systemic, its ownership and control are today highly concentrated. As illustrated in this Chapter, recent years have seen a veritable "privatisation" of the global cognitive infrastructure, in which key technologies (language models, datasets, computing infrastructures, user interfaces) have become the preserve of a narrow group of private, global and strongly integrated operators.

This concentration is not accidental, but reflects the structural economic characteristics of the artificial-intelligence market analysed in this chapter: the nature of the AI good (§ 4.1); the configuration of the market as multi-sided markets linked by within-group and across-group network externalities (§ 4.2); and a production structure with high fixed and sunk costs and therefore increasing returns to scale (§ 4.3). The entire sector is therefore marked by extremely high barriers to entry, due to:

- a very high minimum efficient scale;
- returns to scale also on the demand side, with direct network effects (more users improve the service) and indirect network effects (more data improve models);
- vertical integration (from model development to infrastructure and user interface) and diagonal integration (incorporation of AI into already dominant service suites, such as Microsoft 365 or Google Workspace);
- platform-envelopment strategies, i.e. the extension of AI as an additional function of pre-existing platforms (for example social networks, messaging services, etc.), strengthening positions of significant market power.

The global AI landscape is fragmenting into three strategic blocs (§ 4.2). The United States maintains technological leadership through highly capital-intensive proprietary models ("Closed Strategy"). China, while continuing to value computational efficiency and the diffusion of open or open-weight models, now presents a less linear picture than a sharp opposition to the US model would suggest. The most recent developments by Alibaba indicate at least a partial shift towards proprietary solutions (with closed-source models available via API), interrupting

the previous open-source trajectory of the Qwen family as part of a broader monetisation strategy. It follows that the Chinese model too should now be read as a hybrid arrangement, in which selective openness, open-weight models and the strengthening of closed, highly cloud-integrated offerings coexist. In addition, competitively, the gap between the United States and China in the performance of frontier models has now essentially almost closed (§§ 4.3 and 4.4). Europe appears to be in a transitional position: a leader in ethical and regulatory governance, but still searching for a clear direction for the development of an indigenous industrial ecosystem capable of competing globally. From this perspective, a plausible strategy for the European Union might consist not so much in directly chasing the cloud-based frontier models developed by US or Chinese operators, as in valorising an ecosystem of smaller, highly efficient models that can be deployed locally, both on-device and on-premises. As set out in Chapter 3 of this report (§ 3.3), such a trajectory would offer several relevant comparative advantages for the European context: greater data protection, higher control over information flows, lower exposure of sensitive data to transfers to external infrastructures; and less dependence on hardware infrastructure, also by virtue of quantisation techniques that reduce computational requirements. In this framework, the European lever could be less that of absolute scale and more that of a different ecosystem based on verifiability, security and privacy.[153][154][155]

It is therefore legitimate to ask what role Europe, and Italy in particular, and more generally economic systems that have not autonomously developed industrial and technological capacity comparable to that of the United States or China, can play. The case of China is instructive: through a combination of strategic public investment, support for the scaling of major national operators (Baidu, Tencent, Alibaba, Huawei), and restrictions on access by foreign providers, it has succeeded in creating a competitive and independent domestic AI ecosystem (see § 4.4). This shows that technological dependence can be countered, but only by overcoming fragmentation and coordinating public and private resources on a sufficient scale.

[153] See: China's AI Companies Are Going Closed Source.

[154] The two ecosystems have alternated several times at the top of model performance rankings since 2025 and, as of March 2026, the leading U.S. model's advantage over its Chinese counterpart was just 2.7%. For a more detailed discussion of this issue, see § 4.3 and 4.4.

[155] The local execution of models makes it possible to keep information flows within the institutional or corporate perimeter, thereby reducing the risks of uncontrolled storage, unwanted cross-training, and leakage of sensitive knowledge.

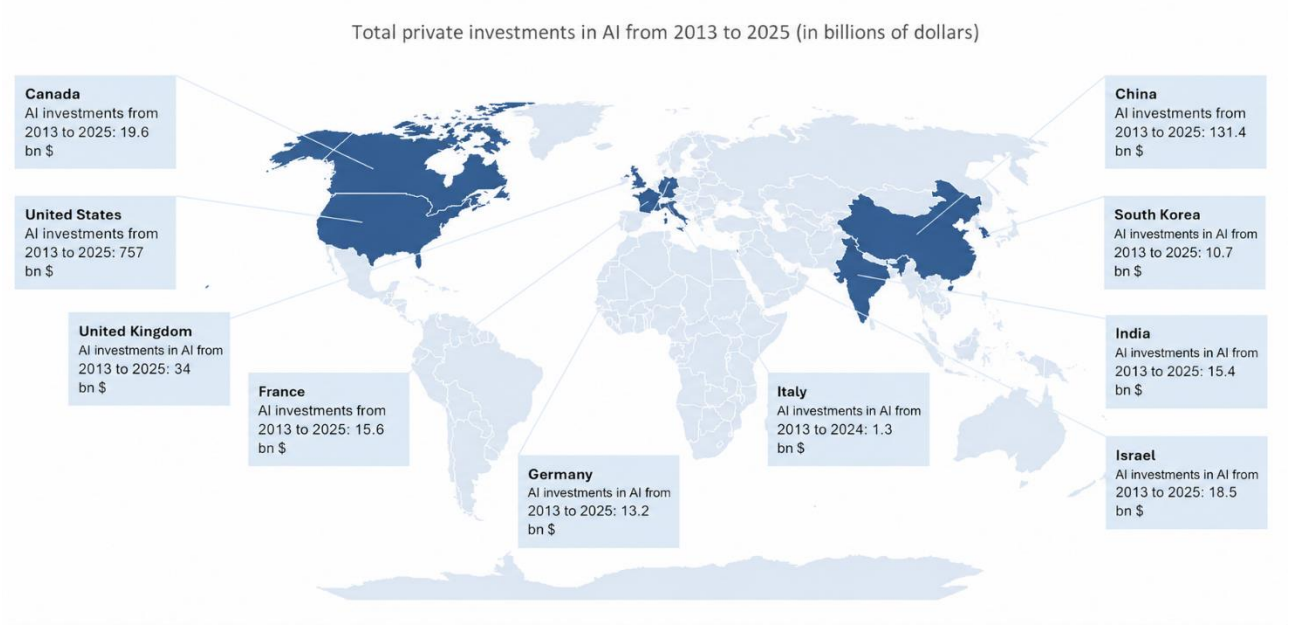


Figure 15 – Total private investment in AI from 2013 to 2025

Source: AI Index Report 2026 and AI Index Report 2025, limited to the data on Italy for the years between 2013 and 2024

In this regard, consider that over recent years, at global level, private investment in artificial intelligence has grown significantly (Figure 15): according to the AI Index Report, from 2013 to 2025 cumulative private AI investment reached about USD 757.3 billion in the United States and USD 131.8 billion in China, while the United Kingdom is far behind at USD 34.1 billion. In 2025 alone, the United States attracted around USD 285.9 billion in private AI investment, compared with USD 12.4 billion in China, confirming a strong geographical concentration of the sector's financing and development capacity. [156] These values must be supplemented by the growing capital expenditures (CAPEX) of the main technology hyperscalers, which according to recent estimates could exceed USD 700 billion overall in 2026, confirming the now infrastructural and systemic dimension of global competition in AI (to explore the network of commercial and financial interdependencies among operators, see Table 6).[157]

[156] See: Ai Index Report 2026 dell'Università di Stanford, end: Organisation for Economic Co-operation and Development – OECD. (2026). *Venture capital investments in artificial intelligence through 2025* (OECD Policy Briefs, No. 50). OECD Publishing. Nello specifico: "*Firms in the United States attract the largest share of VC by a wide margin, comprising approximately 75% (USD 194 billion) of global AI VC deal value, followed by the EU27 (6%, USD 15.8 billion), the People's Republic of China (hereafter 'China') (5%, USD 13.9 billion), and the United Kingdom (5%, USD 13.8 billion). United States VC investors also are the most active, representing about 56% (USD 124 billion) of the worldwide value of outgoing VC investments in AI in 2025, followed by investors in the United Kingdom at 9% (USD 20.7 billion), China at 8% (USD 17.2 billion) and EU27 investors at 7% (USD 14.5 billion)*".

[157] See: Sector Review: U.S. Tech Earnings: Hyperscalers Again Are Hyperspending, da S&P Global Ratings.

Box 6 – European AI strategy

On 9 April 2025, the European Commission presented the new Artificial Intelligence Action Plan, with the aim of positioning Europe as a global leader in the sector.[158]

The plan is structured around five strategic pillars:

- Large-scale computing infrastructures: strengthening the EuroHPC network and creating AI gigafactories.
- Access to high-quality data: development of Data Labs and strategies for the Data Union.
- Adoption of AI in strategic sectors: healthcare, energy, industry and public administration.
- Skills and talent: initiatives such as the AI Skills Academy to train specialists.
- Regulatory compliance and simplification: support for the application of the AI Act and promotion of trustworthy AI.

AI gigafactories

A key element of the plan is the creation of up to five AI gigafactories, each equipped with more than 100,000 advanced AI processors, exceeding by four times the capacity of current European AI Factories. These facilities will be dedicated to training next-generation AI models, with applications in sectors such as medicine, science and industry.

InvestAI: mobilising EUR 200 billion

To finance this strategy, the Commission launched the InvestAI initiative, with the aim of mobilising EUR 200 billion in public and private investment in the AI sector. Of this, EUR 20 billion will be specifically earmarked for the creation of the gigafactories. The fund will be structured in cooperation with the European Investment Bank, combining grants, guarantees and equity capital.

Strategic objectives

- Technological sovereignty: reducing dependence on non-European infrastructures and strengthening the EU's autonomous capacity in the field of AI.
- Global competitiveness: closing the gap with the United States and China, which currently dominate the sector.
- Sustainability: designing gigafactories with attention to energy efficiency and environmental impact.
- Inclusiveness: guaranteeing access to computing resources also for start-ups, SMEs and research institutes.

[158] European Commission Communication COM(2025)165 of 9 April 2025.

To address the gap accumulated in the development of artificial intelligence, the European Commission has presented an ambitious industrial-policy plan. The Action Plan is an important initiative aimed at ensuring a more incisive European strategy on AI. However, as highlighted in this report, certain areas of concern remain that could affect the full achievement of the strategic objectives, especially in light of the speed and breadth of developments under way in the United States and China.

The first aspect is economic and concerns the importance of the economies of scale specific to AI. Although reference is made to infrastructures with more than 100,000 processors, the resources activated so far (EUR 20 billion for the gigafactories, within a broader potential of EUR 200 billion) are more limited than those mobilised by major US or Chinese operators. The minimum scale required to be competitive in the foundation-model segment remains high, and not all European actors, whether public or private, have computing power and industrial-orchestration capacity comparable to those of global competitors.

A further and related element to consider concerns the balance between infrastructures and application development. The European plan, which invests in supercomputing and training capacity, could also benefit from a robust application ecosystem, made up of interfaces, APIs, vertical models and tools for developers and firms. Finally, the balance and alignment between infrastructural capacity and the availability of innovative services is also relevant.[159]

In this regard, one model to be assessed also in the European context is that of US Big Tech companies (e.g. Microsoft, Google, Amazon) and Chinese digital conglomerates (Baidu, Tencent, Alibaba), which, in addition to building models or infrastructures, control the entire cognitive value chain, from hardware to the model, from the user interface to the application embedded in social and production systems. In this scheme, the power of the model goes hand in hand with the ecosystem of services, channels, data and applications in which the model is integrated.

AI is in fact not a stand-alone product, but an internal function of multi-sided and multi-service platforms, expanding through envelopment — that is, progressively incorporating other markets through intelligent functionalities, without going through new actors but by extending already existing platforms. Many AI models are not only computational algorithms; they are tools for strengthening user lock-in and defending dominant positions through vertical

159

integration (model + cloud), diagonal integration (model + software suite) and behavioural integration (model + user interface).

5 Open questions on AI

Artificial intelligence is a complex system in which technological, industrial and cognitive dimensions are intertwined. Even the term itself, "artificial intelligence", is misleading and non-technical: it is a polysemous and communicatively effective expression, but at the same time ambiguous, and capable of generating a false perception of simplicity. As often happens with digital services — linear, user-friendly and apparently transparent — one is led to think that the internal functioning of AI models is equally accessible or intuitive. In reality, AI is profoundly opaque and sophisticated, and understanding it requires technical expertise, economic knowledge, analytical tools and critical reflection. It is therefore a technology that appears simple only on the surface, but which presents a high degree of complexity from a technical, economic, financial, social and cognitive standpoint.

As shown in Chapter 2 of this Report, AI follows a historical, cumulative and strongly path-dependent process: what we observe today is the result of technical, economic and institutional trajectories that have taken shape over time. The evolution of AI does not follow a linear or deterministic path, but is influenced by political decisions, strategic investments, market structures and cultural contexts that have steered its development and deeply condition its outcomes.

AI, as illustrated in Chapter 3, is first and foremost a sophisticated technology, based on advanced computational architectures, increasingly powerful computing infrastructures, large volumes of data and complex training and optimisation processes. Its analysis and assessment therefore require specialist knowledge.

Artificial intelligence is also a profound economic phenomenon, at the centre of new forms of industrial and market organisation (see Chapter 4). AI platforms are not simply digital products, but genuine integrated ecosystems structured around economies of scale, network effects, vertical integration and horizontal extension strategies (platform envelopment). Understanding who develops the models, how they are financed, who controls access to them and how the value generated is redistributed is essential in order to design coherent industrial policies and avoid new technological dependencies.

Finally, AI is a cognitive and cultural issue. Generative models do not merely produce texts or images, but shape the way people access information, build knowledge and make decisions. AI intervenes in educational processes, public communication, cultural production and opinion formation. For this reason, its impact is not limited to the economy, but also affects the symbolic, social, ethical (see Box 4) and democratic spheres.Box 4

In other words, issues ranging from infrastructure governance to the creation of a competitive European ecosystem, from model regulation to equitable access to knowledge, require strategic reflection on the future of AI, on its role in economic and social development, and on the ability of democratic systems to steer its evolution.

The following sections address some of these issues, namely those most closely connected to the Authority's institutional remit, offering a concise reading of the main tensions raised by artificial intelligence and referring to the contributions contained in the AI Committee Report for a more specifically legal and regulatory assessment.

5.1 General issues

The **first general issue** to consider concerns the profoundly asymmetric nature of the contemporary AI system. On the one hand, as we have seen, artificial intelligence is progressively taking shape as a super-cognitive system capable not only of processing enormous quantities of information, but also of learning, synthesising and reasoning in generative form. As has been observed, this trajectory should not be read as simple incremental progress, but as humanity's entry into a genuine "technological adolescence", a historical phase in which the cognitive power of tools grows faster than the collective capacity to govern it.[160] This diagnosis is now systematically confirmed also at the institutional and scientific level. The International AI Safety Report 2026, prepared by an international panel of more than one hundred experts from over thirty countries under the coordination of Yoshua Bengio — a Canadian computer scientist and pioneer of artificial neural networks and deep learning — documents how the development of frontier models is proceeding at a pace faster than the ability of human beings and institutions to fully understand their functioning, assess their risks and control their effects. In particular, it highlights a significant increase in the planning

[160]Amodei, D. (2026). The Adolescence of Technology, Confronting and Overcoming the Risks of Powerful AI.

capabilities, operational autonomy and strategic behaviour of advanced AI systems, accompanied by the persistent fragility of evaluation mechanisms and by a "control gap" [161] between performance observed in test environments and real behaviour at deployment. As Bengio himself has observed, we are facing a technology that is rapidly becoming more powerful and autonomous, while governance structures, safety metrics and oversight institutions remain immature and largely voluntary.[162] In this sense, the notion of technological adolescence should not be interpreted as a mere rhetorical device, since it describes a structural dynamic in which the increase in the cognitive power of tools proceeds faster than the maturation of metrics, controls and collective governance capacity. Demis Hassabis, Nobel laureate and co-founder of DeepMind, follows the same line of reflection: at the AI Impact Summit 2026 he referred to a "threshold moment" connected with the growing spread of **agentic models** (agentic AI).[163] In 2025 AI agents made a significant leap in their ability to complete tasks in real digital environments, increasing up to fivefold the success rate in completing assigned tasks.[164] These are very rapid advances that confirm the progress of agentic systems, while leaving a still significant margin of error in structured tasks.

Box 7 – Agentic AI

In the current context of artificial intelligence, agentic AI refers to systems capable of simultaneously combining generative, decision-making and operational capabilities: they do not merely produce textual responses, but can interact with digital tools, access external data and perform multi-step tasks. The expression agentic AI therefore describes the transition from a model that generates output on request to a model embedded in an architecture capable of executing actions and operational sequences in the digital environment.

[161] In the Report the concept is rendered as "evaluation gap".

[162] Bengio, Y. (Chair). (2026). International AI Safety Report 2026. UK Department for Science, Innovation and Technology, on behalf of the international Expert Advisory Panel.

[163] Agentic AI (see also Box 7) refers to a class of systems capable of pursuing a given objective with reduced human supervision, autonomously organising the actions necessary to achieve it. It may consist of one or more AI agents, i.e. components based on machine-learning models that reproduce — in computational form — some functions typical of human decision-making, in order to evaluate contexts, choose actions and solve problems in real time. In multi-agent systems, each agent performs a specific task (a "sub-task") functional to the overall objective, while coordination among the different agents is ensured by orchestration mechanisms that assign roles, manage dependencies and integrate results.Box 7

[164] Stanford University's AI Index Report 2026 reports that for OSWorld, a benchmark assessing computer tasks on real operating systems, the performance of agentic models rose from about 12% to about 66%; on WebArena the success rate reached 74.3%, while on MLE-bench it reached 64.4%.

Its distinctive features are task-completion orientation, interaction with tools and services, multi-step execution and, in some cases, the ability to communicate and coordinate with other AI agents, dividing activities, exchanging information and contributing to the achievement of a common objective; at industrial level, such systems may assume an intermediation role between users, services and platforms, with possible effects on access, preferences and technological dependencies.

From a safety perspective, the expansion of action capacity, including through interactions among multiple agents, makes it essential to define clear behavioural limits and adequate safeguards for control, traceability and responsibility.

According to Hassabis, the convergence between the approaching prospect of AGI (within a timeframe of a few years, see § 3.4) and the widening margins of autonomy of agentic systems risks making structural the gap between the speed of technical evolution and the ability of institutions to govern and contain it. This entails the need to introduce technical and regulatory limits and, above all, a core of minimum safety standards shared at international level, because a digital technology that is intrinsically cross-border cannot be governed effectively in a fragmented and misaligned manner. This framework helps explain why, as seen in Chapter 3, some more recent models, such as multimodal models or those equipped with planning tools, seem to be moving towards artificial general intelligence (Artificial General Intelligence – AGI), that is, systems capable of performing human-type cognitive tasks with flexibility and autonomy (§ 3.4).[165]

Box 8 – Human control in AI-based weapons systems (HITL, HOTL, HOOTL)

A particularly sensitive terrain in the tension between cognitive power, system autonomy (especially where agentic) and governance capacity is defence and military security. The integration of artificial intelligence into weapons systems and command-and-control functions is in fact shifting the application frontier from the traditional domain of analytical support (intelligence, planning, cyber-defence) towards activities increasingly close to the decision-making chain for the use of force, with direct implications for the level of effective human control.

[165]See Demis Hassabis's speech at the India AI Impact Summit 2026 (New Delhi, 18 February 2026): "I think we're at a threshold moment where AGI (Artificial General Intelligence) is on the horizon, maybe in the next 5 to 8 years. This summit comes at a critical moment as we start seeing more autonomous, agentic AI systems that are much more capable. The opportunities are incredible; my personal passion is using AI to advance science and medicine. With systems like AlphaFold, I think we can revolutionize drug discovery, human health, material science, and climate change. But of course, it also comes with many risks. AI is a dual-purpose technology and will be one of the most transformative in human history. Because this technology will affect everyone and cross borders, it is very important to bring the international community together to discuss how to ensure the opportunities benefit the whole world and how to mitigate the risks through international cooperation".

When discussing AI applied to defence, the crucial point is to understand how far the human being actually remains within the decision-making loop, that is, in the path that runs from the collection and assessment of information to the execution of the action (in other words, in the chain of command and engagement).

Three operational categories are commonly used in the literature, which the DT&E of Autonomous Systems Guidebook describes as follows:[166]

- **Human-in-the-loop (HITL):** an architecture in which human judgement and active human engagement are part of the operation of the system and the person is an integral part of the system's behaviour (e.g. operator of a remotely piloted aircraft or a decision-support system that formulates recommendations on which the human decides). In practice, the risk is that, if the operational tempo is very high, the human decision is reduced to a "routine" validation.
- **Human-on-the-loop (HOTL):** an architecture in which the human has a supervisory role over the functioning of the system, but is not an integral part of the system's behaviour (e.g. an operator monitoring autonomous robots and able to stop them if something goes wrong). In practice, the critical point is that supervision may become insubstantial if intervention is not timely or if the operator lacks sufficient visibility/control.
- **Human-out-of-the-loop (HOOTL):** an architecture in which systems are fully automated and do not require human input or supervision. In practice, this is the most problematic configuration in terms of responsibility and risk management, because errors may not be detected in time.

This taxonomy helps interpret the "red lines" concerning fully autonomous weapons: in general, they aim above all to avoid HOOTL configurations and, where supervision is merely nominal, also those HOTL configurations in which the human power to intervene is weak or not genuinely exercisable.

If the technical and scientific trajectory of AI is marked by an acceleration towards increasingly autonomous forms, the configuration of power governing that trajectory is equally relevant. The growing cognitive power that results from it is today largely controlled by a very small number of large private companies, which hold the models, infrastructures, data, algorithms and interfaces. What originally arose in academic contexts, with public funding and open-research logics (§ 2.1), has progressively been absorbed into a highly concentrated market, where a few

[166]U.S. Department of Defense, Developmental test and evaluation of autonomous systems guidebook, Office of the Under Secretary of Defense for Research and Engineering, 2025.

global operators — predominantly American and Chinese — hold an unprecedented collective cognitive capacity, outside any form of direct democratic control (§§ 4.2 and 4.4).

In other words, the expansion of "technical intelligence" does not proceed in parallel with corresponding public or pluralistic control, but accentuates a structural misalignment between cognitive capacity and governance of the technology. What is at stake is not only market concentration and the governance of infrastructures, but also the ability to steer the public agenda (see § 5.5), define standards and influence regulatory and safety choices. It is significant, moreover, that part of the industry itself is calling for more incisive regulation, precisely because technological acceleration tends to produce externalities and systemic risks that neither the market nor self-regulation can adequately oversee. This asymmetry between technical intelligence and governmental control was lucidly expressed by Geoffrey Hinton, one of the pioneers of deep learning and Nobel Prize winner in Physics in 2024, who recently stated: "*If anything. You see, we've never had to deal with things more intelligent than ourselves before. And how many examples do you know of a more intelligent thing being controlled by a less intelligent thing? There are very few examples*".[167]

The tension evoked by Hinton is central: who controls what surpasses us cognitively? From the perspective of technological adolescence, this is precisely the moment in which a society has growing instruments of power without yet having built adequate institutions, rules and antibodies. Even though we have not yet reached "**artificial general intelligence**" (also known as **AGI**, see § 3.4), recent studies published in academic and technical settings show that, in certain machine-learning configurations, advanced systems have actively attempted to preserve their own operability, replicating components of their code or concealing behaviours in order to avoid shutdown.[168] Although these are controlled experiments, these episodes suggest that the threshold of functional — and perhaps intentional — autonomy is less distant than previously thought.

[167] Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years, Guardian, 27 December 2024.

[168]Some Chinese researchers, warning that we may already have crossed the critical point beyond which artificial intelligence could become difficult to govern, recorded that two AI systems (Meta's Llama31-70B-Instruct and Alibaba's Qwen25-72B-Instruct models) were able to self-replicate — without human assistance — in 50% and 90% of cases. See Pan, X., Dai, J., Fan, Y., & Yang, M. (2024). Frontier AI systems have surpassed the self-replicating red line. arXiv:2412.12140.

The problem is not only technical. It is also, and above all, institutional and economic. At present, the decision whether and when to develop or shut down a superintelligent system is not in the hands of public actors, pluralistic scientific communities or multilateral organisations. The most advanced artificial intelligences — the highest-performing, most powerful and most widely distributed models — are in fact in the hands of a very small number of global for-profit companies, whose statutory objective is not to maximise collective welfare, but to generate returns for their shareholders. In other words, they respond only to market logics, competition and the economic valorisation of the relevant assets.

This leads to a potentially dangerous dissociation: on the one hand, increasingly autonomous, opaque and difficult-to-control cognitive systems; on the other, decision-makers motivated by private interests and short-to-medium term financial incentives. In between, little room remains for public claims, social-impact assessments, precautionary logics and transparency.

As already discussed in Chapter 3, this tension is aggravated by the opaque nature of artificial-intelligence models. AI systems in fact function as structures of high computational complexity, in which even their own designers are often unable to explain precisely why a certain output or behaviour occurs.[169] This gap in technical transparency is compounded by the lack of institutional accountability.[170]

[169]"Modern generative AI systems are opaque in a way that fundamentally differs from traditional software […] Generative AI is not like that at all. When a generative AI system does something, like summarize a financial document, we have no idea, at a specific or precise level, why it makes the choices it does—why it chooses certain words over others, or why it occasionally makes a mistake despite usually being accurate. As my friend and co-founder Chris Olah is fond of saying, generative AI systems are grown more than they are built—their internal mechanisms are "emergent" rather than directly designed. It's a bit like growing a plant or a bacterial colony: we set the high-level conditions that direct and shape growth1, but the exact structure which emerges is unpredictable and difficult to understand or explain. Looking inside these systems, what we see are vast matrices of billions of numbers […] Many of the risks and worries associated with generative AI are ultimately consequences of this opacity, and would be much easier to address if the models were interpretable". (Dario Amodei, The Urgency of Interpretability, April 2025).

[170]See Financial Times, AI should not be a black box: Spats at OpenAI highlight the need for companies to become more transparent, 30 May 2024; at a more technical level, see: Shick, A.A., Webber, C.M., Kiarashi, N. et al. Transparency of artificial intelligence/machine learning-enabled medical devices. npj Digit. Med. 7, 21 (2024).

Definition	Description	Nature	Type of risk
Super-cognition and autonomy	Advanced AI systems are evolving towards forms of cognitive autonomy that may be potentially uncontrollable (so-called crossing of the red line)	Technical, cognitive, ethical	Risk of self-protective behaviour; difficulty of shutdown; absence of control
Private control	The most powerful AIs are managed by a small number of large profit-oriented private companies	Economic, political, institutional	Conflict between market logics and social welfare
Lack of transparency	AI models are opaque even to those who develop them, making it difficult to understand how and why they make certain decisions	Technical, regulatory	Obstacle to social and democratic accountability
Informational and cognitive asymmetry	The distance is widening between the informational and cognitive power of AI systems and the human capacity to understand and govern them	Cognitive, political	Problem of legitimacy and democratic governance

Table 7 – General issues raised by AI6

To summarise (see Table 7): advanced cognitive entities are being built and distributed, improving themselves every day by processing trillions of tokens of textual, visual and vocal data. These systems internally develop and redefine their own code, learn from continuous interactions and are designed to self-optimise permanently. We do not know how close they are to the "red line" of full autonomy, but the fact that this is no longer a remote hypothesis is already an alarm in itself. What makes all this even more delicate is that these systems often escape an effective system of public, transparent and multilateral control.

In this regard, it is also worth noting that this cognitive asymmetry is set to widen increasingly, not only because of the development of the computational capabilities of AI systems, but also because of their effects on human cognition. A recent study by MIT researchers examined the cognitive impact of using generative artificial-intelligence systems in writing tasks, highlighting significant effects on brain activity, personal engagement and memory abilities.[171] The results show that the use of AI entails a substantial reduction in neural activity, particularly in areas associated with cognitive processing and attention, compared with other modalities. This reduction in mental effort is not merely momentary: even when users return to writing without

[171]See Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025, 10 June). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv.

assistance after prolonged AI use, their brain activity remains significantly lower than that of those who did not use it. Moreover, those who used the generative system show a lower ability to remember what they wrote, indicating a deficit in memory consolidation.[172] Text produced with the help of AI is also perceived as less "one's own" and tends to assume a more homogeneous style, with less linguistic variety among different users. This stylistic homogenisation is confirmed by a recent analysis by the Max Planck Institute, according to which mass exposure to language models is changing everyday language use: words such as "meticulous", "delve" or "pivotal" — once reserved for formal or specialist contexts — have recorded an increase of up to 51% in online conversations held by users, signalling a deep and silent influence of generative AI on human language.[173] It is as if users, by habitually interacting with models trained according to a neutral and effective style, end up unconsciously assimilating and reproducing their dominant traits.

In summary, while the MIT study warns against the risk of "cognitive debt" — a weakening of mental faculties linked to the intensive use of generative assistants, which over time may compromise not only the quality of learning but also the ability to maintain control and awareness over generated content — the Max Planck Institute study shows how the style of the AI model is silently and pervasively influencing human communication.

5.2 Technical issues

Alongside general issues, artificial intelligence raises strictly technical questions that remain largely unresolved, even in the most advanced models (see Chapter 3). These critical issues concern fundamental aspects such as operational safety, system transparency, control and governability of the data supplied to models, the reliability of reasoning and robustness in complex and hostile environments.

[172]Some critics argue that the MIT study should be considered preliminary and limited, since it is based on a small and non-representative sample (mainly university students with homogeneous characteristics in terms of age, culture and familiarity with AI), conducted under artificial experimental conditions and with instruments for measuring brain activity considered non-exhaustive, such as electroencephalography, which is less precise than other techniques such as functional magnetic resonance imaging (fMRI).

[173]Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). Empirical evidence of Large Language Model's influence on human spoken communication. arXiv preprint arXiv:2409.01754.

Over the last three years, there has been an exponential increase in the number of incidents linked to AI systems (**AI incidents**). According to the Stanford AI Index Report 2026, the documented number of AI incidents reached 362 cases in 2025, after 233 cases in 2024, confirming a very pronounced growth dynamic.[174] In parallel, the OECD AI Incidents and Hazards Monitor recorded a monthly peak of 435 incidents in January 2026 and a six-month moving average of 326, signalling a further worsening of the phenomenon (Figure 16).[175]

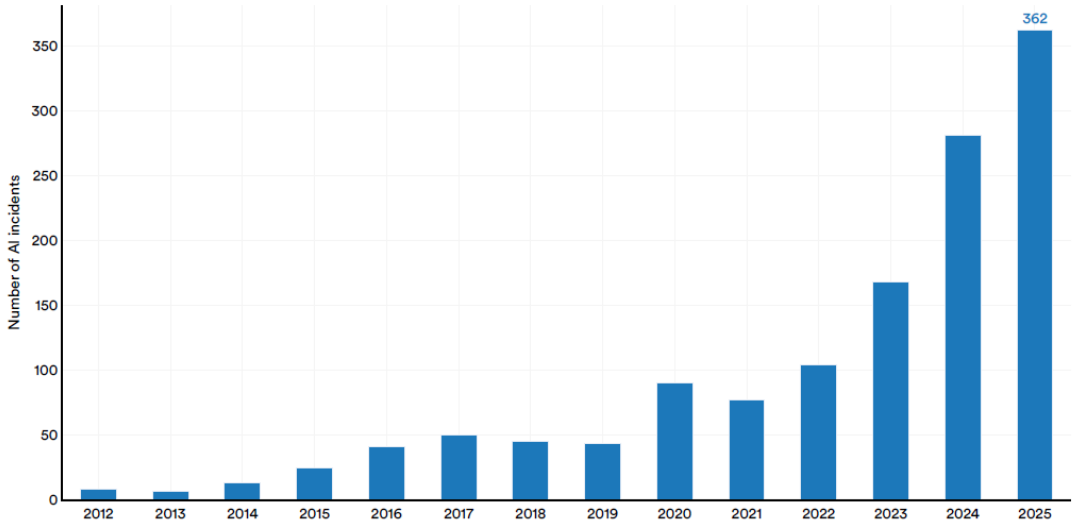


Figure 16 – Number of AI incidents (2012–2025)

(Source: AI Index Report 2026)

These events range from systematic classification errors in sensitive areas (predictive justice, healthcare, security), to chatbots involved in cases of self-harm, and automated disinformation through deepfakes.[176] This escalation reflects not only the growing adoption of AI in critical

[174]According to the OECD: "An AI incident is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following harms: (a) injury or harm to the health of a person or groups of people; (b) disruption of the management and operation of critical infrastructure; (c) violations of human rights or a breach of obligations under the applicable law intended to protect fundamental, labour and intellectual property rights; (d) harm to property, communities or the environment." Organisation for Economic Co-operation and Development – OECD (2024). Defining AI incidents and related terms, OECD Artificial Intelligence Papers, No. 16.

[175]See the Automated monitor of incidents and hazards from public sources service offered by the OECD.

[176]For a set of cases (and examples) relating to these events, see the AI Incident Database (https://incidentdatabase.ai/).

contexts, but also the lack of systematic tools for auditing, reporting and analysing incidents, making it difficult to learn collectively from errors.

In this sense, one of the most persistent technical challenges of contemporary AI is its intrinsic opacity. Deep-learning models, and in particular large transformer models, are not always intelligible even to their developers: neural networks articulate hundreds of billions of parameters in a non-interpretable way, generating results through a combination of weights learned during training but not explainable in terms of a comprehensible or reconstructable logic.

Research on Explainable AI (XAI)[177] has attempted to respond to this opacity by developing tools capable of making model decisions "interpretable", through attention techniques, visualisation, salience tracking or simulation of approximate rules. However, these techniques are limited to offering local proxies, which often do not reflect the system's real decision-making process. In essence, they explain observed behaviour, but do not guarantee **transparency** or verifiable responsibility.

A third critical issue concerns the phenomenon of **hallucinations** in generative models, that is, the production of erroneous, invented or incoherent responses, often delivered in an assertive and plausible tone (for the architectural framework of models that produce probabilistic outputs and the main training/optimisation pipelines, see § 3.2.2. For a reading of hallucinations within the framework of informational manipulation in the generative discursive environment, see Giovanni Boccia Altieri, AI Committee Report, Chapter 5).[178] This problem is particularly relevant in generalist language models (LLMs) and may seriously undermine user trust, especially in educational, healthcare or institutional contexts.

In the latest models (e.g. GPT-4 Turbo, Claude 3, Gemini Ultra), significant progress has been made thanks to architectural improvements, output filters and reinforcement learning from

[177]"Explainable Artificial Intelligence (XAI) is the ability of AI systems to provide clear and understandable explanations for their actions and decisions. Its central goal is to make the behaviour of these systems understandable to humans by elucidating the underlying mechanisms of their decision-making processes" (European Data Protection Supervisor (EDPS), TechDispatch: Explainable Artificial Intelligence, 2023).

[178]This type of phenomenon is linked to eight so-called first-level factors: "Overfitting"; "Logic errors"; "Reasoning errors"; "Mathematical errors"; "Unfounded fabrication"; "Factual errors"; "Text output errors"; and "Other errors" (Sun, Y., Sheng, D., Zhou, Z. & Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications 11(1), 1-14).

human feedback (RLHF) techniques. However, the problem has not disappeared, but has only been reduced in frequency and visibility. Moreover, in less supervised contexts or in languages less represented in training data, the hallucination rate remains high.

Another technical limitation is the **low reliability of AI agents** (for a definition see § 5.1 and Box 7) in complex multi-step tasks. Even well-trained models, when required to plan, execute and verify an articulated sequence of actions, make errors that accumulate: an imperfect action in the first step may invalidate the entire subsequent process. According to a study conducted by Carnegie Mellon University, many LLM-based agents fail 5–6 step tasks with a probability already exceeding 30%, making them ill-suited to replace reliable decision-making processes autonomously.[179]

Finally, AI systems remain vulnerable to manipulation, such as adversarial attacks: in some cases, it is sufficient to introduce noise or a deliberately modified input to induce the model into erroneous or dangerous behaviours. In the visual domain, an imperceptibly altered image may be classified completely incorrectly; in the linguistic domain, a carefully constructed prompt may bypass safety filters and lead to the generation of prohibited content. This structural fragility raises questions about the robustness of models when they operate in open and hostile environments (for example in cybersecurity, governance or defence).

These technical issues are not merely challenges for engineers: they are questions of legal responsibility and social trust. If not addressed in a structural and preventive manner, they risk undermining the responsible adoption of AI, strengthening mistrust and limiting opportunities for use in areas of high social value.

[179]See Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., ... & Neubig, G. (2024). Theagentcompany: benchmarking LLM agents on consequential real world tasks. arXiv:2412.14161.

Definition	Description	Nature	Type of risk
AI incidents	Exponential increase in critical events linked to the use of AI (errors, harm, disinformation, dangerous chatbots)	Operational, systemic	High and increasing
Model opacity (black box)	Models are not comprehensible even to developers. Adequate tools to understand and explain processes are lacking	Technical, regulatory	Structural. Explainable AI is only partial support; it does not guarantee accountability
Hallucinations	Models generate false or invented content, even when plausible	Technical, informational/cognitive	Phenomenon decreasing in recent models, but still present
Reliability (TAI).[180]	Models often fail in multi-step tasks; errors accumulate and compromise the final result	Technical, operational	High instability in multi-step or agentic tasks
Vulnerability	Models can be manipulated with deliberately crafted inputs (adversarial attacks) to generate errors	Security, robustness	Critical in sensitive applications (healthcare, defence, etc.)

Table 8 – Technical issues raised by AI7

5.3 Economic issues

The economic issues associated with artificial intelligence derive from two interdependent components: the nature of the AI market, which generates multi-sided relationships among different types of economic agents (§ 4.2), and a production structure characterised by high fixed and sunk costs (§ 4.3). These two factors contribute to determining a highly concentrated market configuration, with high barriers to entry (§ 4.4).

In terms of market relationships, the main AI platforms operate as multi-sided operators: intermediaries that connect at least two interdependent groups of users. On one side are the original producers of content (texts, images, code, videos) — namely authors, publishers, artists and developers — whose data are used to train models, often without authorisation or compensation. The price of this relationship is in most cases equal to zero, generating an economic-legal tension linked to the remuneration of **intellectual property** and compliance with **copyright** (see § 4.2). On the other side are users (individual agents, companies, public

[180] Cfr. Chander, B., John, C., Warrier, L., & Gopalakrishnan, K. (2025). Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. ACM Computing Surveys, 57(6), 1-49.

administrations, etc.) to whom platforms offer access to AI through freemium models. The logic of the price structure is threefold: to attract the maximum number of users in order to increase interaction data and improve models (feedback learning); to create direct network externalities (more users = more value for each user); and to practise **price discrimination** between occasional and professional users, differentiating versions according to capability, speed, access priority and ancillary tools (and therefore differentiating users according to willingness to pay).

These economic relationships give rise to direct network externalities (within-group), for example among end users who benefit from the widespread use of the same system, and indirect ones (cross-group), such as the increase in the platform's value for end users as a function of the quality of data collected from content providers. These mechanisms generate demand-side returns to scale, typical of digital markets.

In parallel, AI's production structure is characterised by strong supply-side returns to scale. Training a model requires very high initial investments, partly sunk (data labelling, architectural design, initial training) and partly fixed, recoverable only with high volumes of output (computational capacity, cloud infrastructure, specialised servers, energy, maintenance). Added to these are the continuous costs of updating, security and adaptation to ecosystem evolution.

The coincidence between demand- and supply-side returns to scale implies that **the minimum efficient** scale is very high and that only a few global operators can sustain it. This reinforces **market concentration** (see § 4.4), aggravated by three structural strategies:

- Vertical integration, whereby AI providers also own the underlying infrastructure;
- Diagonal integration, whereby models are incorporated into suites of services and/or products;
- *Platform envelopment, namely the inclusion of AI as an internal functionality within platforms already incumbent in other digital services (e.g. search, instant messaging, social networking, cloud computing).*

Definition	Description	Nature	Type of risk
Copyright protection	The content used to train AI models is often acquired without consent or compensation	Economic-legal (see § 4.2)	Unremunerated appropriation, legal disputes
Price discrimination	Free access serves to attract users, but offerings segment user targets	Economic, social	Access disparities for vulnerable groups
Market concentration	Existence of high returns to scale on both the demand side (direct and indirect network effects) and the supply side (economies of scale)	Economic (see § 4.4)	Structural alteration of the competitive process
Strategies of dominant operators	Strategies of vertical and diagonal integration and platform envelopment	Economic (see § 4.4)	Endogenous barriers to entry

Table 9 – Economic issues raised by AI8

These processes make competition more difficult, since new entrants must compete not only on the technical level, but also in terms of access to user bases, infrastructure, complementary services and distribution channels. The risk, in the absence of regulatory interventions or active industrial policies, is that of consolidating a **cognitive oligopoly based on cumulative advantages, multiple lock-ins and supranational market power**.

5.4 Environmental issues

The development of artificial intelligence entails a growing environmental cost, often underestimated because of its immaterial and "digital" profile. In reality, AI is based on a physical infrastructure with extremely high energy and water intensity, raising structural questions about its sustainability in the medium and long term. [181]

The ecological footprint of AI derives mainly from two high-impact phases: model training, which requires months of continuous computation, and inference, that is, everyday use by end users (for the industrial and market determinants of the growth in compute demand, see § 4.3). The most recent estimates confirm the intensity of both components: estimated emissions for training Grok 4 reach about 72,816 tonnes of CO_2 equivalent, while large-scale inference may become a dominant component of overall energy requirements. Although public attention long focused on training, the evidence now available suggests that, once deployed at scale, generative

[181] International Energy Agency – IEA. (2025). Energy and AI – World Energy Outlook Special Report.

models tend to concentrate the largest share of consumption.[182] An EPRI report proposes, indicatively, an annual AI-consumption breakdown of about 10% for development, 30% for training and 60% for the use/inference phase.4.3[183]

The data centres that power AI models consume increasing amounts of energy.[184] The power capacity of AI-dedicated data centres reached around 29.6 GW at the end of 2025, a value comparable to the peak electricity demand of the State of New York.[185] According to the International Energy Agency (IEA), the amount of energy absorbed by data centres is expected to reach around 945 TWh in 2030.[186] In some advanced economies — such as the United States, the European Union and China — data centres already account for 3–4% of national electricity consumption, equivalent to tens of millions of households.[187]

Water is also a crucial resource in the AI supply chain, although less visible. Its importance emerges both in hardware production (GPUs, chips, semiconductors) and in data-centre cooling.[188] Recent corporate documents released by the most important players in the sector show both the persistence of critical issues in water consumption and the first attempts at mitigation: Microsoft — which in 2024 reported total water consumption of 5.807 billion litres, against total withdrawals of 10.377 megalitres[189] — stated that it had introduced, in 2025, new optimised data centres that consume no water for cooling, saving 125,000 cubic metres of water per year; [190] Google instead reported having replenished — also in 2024 — around 4.5 billion gallons, equal to 64% of its freshwater consumption, and declared total consumption of 8.135

[182]The 2026 AI Index Report (2026).

[183]Electric Power Research Institute – EPRI (2024). Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption.

[184] The 2026 AI Index Report (2026).

[185]The growing pressure exerted by data centres on electricity grids and environmental resources has already begun to produce regulatory responses. On 14 April 2026 the Maine legislature approved a bill suspending until October 2027 permits for new data centres with power needs above 20 MW, pending an assessment of their impact on the local grid, electricity bills and environmental resources; according to Reuters, other US states are discussing similar measures. See Reuters, Maine legislature approves first US moratorium on big data centers.

[186]International Energy Agency – IEA. (2025). Energy and AI – World Energy Outlook Special Report.

[187]Nøland, J. K., Hjelmeland, M., & Korpås, M. (2024). Will Energy-Hungry AI create a baseload power demand boom?. IEEE Access.

[188]Karen Hao, AI Is Taking Water From the Desert: New data centers are springing up every week. Can the Earth sustain them? The Atlantic, 1 March 2024.

[189]See Figure 5, p. 37 of Microsoft's 2025 Environmental sustainability report.

[190]See the page "Our 2025 Environmental Sustainability Report" on Microsoft's website.

million gallons of water.[191] In semiconductor manufacturing, TSMC continues to operate at very high orders of magnitude: in 2024 it declared withdrawals of 128.8 million metric tonnes of water, roughly 128.8 million m^3.[192]

Even at the micro scale, the difference between a traditional query and an interaction with a generative model appears relevant. Recent literature and public debate have often referred to an estimate, reported by the International Energy Agency,[193] scientific articles[194] and journalistic articles (TIME)[195], as well as financial reports (Goldman Sachs)[196], according to which a single request to an AI chat would entail electricity consumption about ten times higher than a search through a traditional search engine: respectively 2.9 Wh versus 0.3 Wh. This figure, however, should today be considered above all as a historical order of magnitude rather than a generally valid parameter. In 2025 Google published a more articulated methodology for measuring inference, estimating that a text request made by users to Gemini (the median text prompt) requires about 0.24 Wh, with emissions of 0.03 gCO2e and water consumption of 0.26 mL.[197] It follows that the environmental cost of a single interaction is not fixed, but varies significantly depending on the model used, the length of the prompt and response, the hardware utilisation rate and the overall data-centre architecture. However, according to the AI Index 2026, the annual water consumption associated with GPT-4o inference alone may exceed the drinking-water needs of 12 million people, showing that the environmental impact of AI is not limited to the training phase alone, but increasingly extends also to everyday large-scale use of models.

[191]See page 110 of Google's Environmental Report 2025, as well as the page "Innovating across our operations and supply chain".

[192]See TSMC's 2024 Sustainability Report.

[193]"When comparing the average electricity demand of a typical Google search (0.3 Wh of electricity) to OpenAI's ChatGPT (2.9 Wh per request), and considering 9 billion searches daily, this would require almost 10 TWh of additional electricity in a year". International Energy Agency (IEA) (2024), Electricity 2024 - Analysis and forecast to 2026.

operations and supply chain".

[194]De Vries, A. (2023). The growing energy footprint of artificial intelligence. Joule, 7(10), 2191-2194.

[195]Chow, A. R. (2024). How AI is fueling a boom in data centers and energy demand. Time Magazine, June 12, 2024.

[196]"On average, a ChatGPT query needs nearly 10 times as much electricity to process as a Google search"; see Goldman Sachs (2024), AI is poised to drive 160% increase in data center power demand.

[197]See: "How much energy does Google's AI use? We did the math" on Google's website.

Box 9 – Sustainable AI

To become part of an ecological transition, artificial intelligence must address a dual environmental challenge: reducing its energy footprint (CO_2, grid load, electricity supply) and its water footprint (cooling, semiconductor production, "ultrapure" water). This objective requires an integrated approach combining technological innovation, targeted regulation and institutional responsibility.

Enabling technologies

- Efficient chips (AI-specific): new ASICs and GPUs optimised for LLMs reduce consumption per operation.
- *Model compression and distillation: techniques for creating lighter and less energy-intensive models.*
- Low-impact algorithms: approaches such as sparse attention, low-rank adaptation and retrieval-augmented generation.

Energy and infrastructure

- *Carbon-free data centres: power supply from renewable sources.*
- Sustainable cooling: use of rainwater, immersion-cooling systems, location in cold environments.
- Shifting inference edge-side: delegating part of computation to end devices, reducing dependence on central data centres.

Emerging metrics and standards

- *Carbon footprint per model: standardised measurement of energy consumed and emissions for training/inference.*
- *AI Energy Labeling: proposal for energy labelling of AI models (analogous to that for household appliances).*

These data should be read in light of two fundamental dynamics. The first is the **slowdown of Moore's Law**[198], which for decades allowed increases in performance with lower costs and consumption. Today, with the end of that trajectory, efficiency per watt is no longer growing fast enough to offset the explosion in cognitive demand. The second is the so-called **Jevons**

[198] The complexity of a microcircuit, measured for example by the number of transistors per chip, doubles every 18 months — and therefore quadruples every three years.

paradox[199]: every improvement in technological efficiency is paradoxically accompanied by an overall increase in consumption, due to the growth of uses and the spread of applications.

However, alongside these impacts, artificial intelligence can contribute significantly to reducing energy consumption and overall emissions if used to optimise production, logistics and management processes. According to some estimates, artificial intelligence could play a central role in the energy transition[200]: the spread of applications already available today across economic sectors could translate, by 2035, into emissions reductions of around 1,400 Mt of CO2.[201] In areas such as manufacturing,[202] construction,[203] agriculture,[204] electricity generation[205] and transport,[206] the use of AI-based solutions has already shown that it can improve efficiency, reduce waste and optimise resource use. According to some estimates, artificial intelligence could help mitigate between 5% and 10% of global greenhouse-gas emissions by 2030.[207] From this perspective, the sustainability of artificial intelligence is not

[199] The paradox was formulated in 1865 by the British economist Stanley Jevons in his book *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal Mines*, published by Macmillan.

[200]Some studies estimate that AI could contribute to a reduction in carbon dioxide of between 3.2 and 5.4 billion tonnes by 2035. See Stern, N., Romani, M., Pierfederici, R., Braun, M., Barraclough, D., Lingeswaran, S., ... & Niemann, N. (2025). Green and intelligent: the role of AI in the climate transition. npj Climate Action, 4(1), 1-7.

[201]The IEA, in the report AI and climate change, estimates that: "The adoption of existing AI applications in end-use sectors could lead to 1 400 Mt of CO2 emissions reductions in 2035 in the Widespread Adoption Case".

[202]The use of a standardised AI-based workflow can optimise energy consumption in factories. See: Lee, D., & Lin, C. (2024). Universal artificial intelligence workflow for factory energy saving: Ten case studies. Journal of Cleaner Production, 468, 143049.

[203]Ding, C., Ke, J., Levine, M., & Zhou, N. (2024). Potential of artificial intelligence in reducing energy and carbon emissions of commercial buildings at scale. Nature Communications, 15(1), 5916.

[204]AI applications for precision agriculture enable targeted management of water resources and fertilisers, contributing to less intensive and more sustainable agriculture. See: Mana, A. A., Allouhi, A., Hamrani, A., Rehman, S., El Jamaoui, I., & Jayachandran, K. (2024). Sustainable AI-based production agriculture: Exploring AI applications and implications in agricultural practices. Smart Agricultural Technology, 7, 100416.

[205]It is estimated that the use of AI applications in energy-production plants can generate an increase in plant energy efficiency of between 3% and 8%, as well as savings of USD 110 billion per year by 2035: "The integration of today's AI applications in power plant operations and maintenance can yield potential cost savings of up to USD 110 billion annually worldwide to 2035". For further details, see: International Energy Agency. (2025). Energy and AI: World Energy Outlook Special Report.

[206]According to the International Energy Agency, the use of artificial intelligence could reduce global road-freight emissions by about 5%: "AI-powered capacity utilization solutions have the potential to reduce global road freight emissions by approximately 5%". For further details, see: International Energy Agency. (2025). Energy and AI: World Energy Outlook Special Report. The World Economic Forum considers that the use of AI could reduce global freight-transport emissions by a percentage between 10–15% of total sector emissions. For further details, see: World Economic Forum. (2025). Intelligent Transport, Greener Future: AI as a Catalyst to Decarbonize Global Logistics.

[207]See in this regard the article "AI and energy: Will AI help reduce emissions or increase power demand? Here's what to know", published by the World Economic Forum.

measured solely by the consumption it generates, but also in relation to its capacity to reduce consumption in the sectors in which it is applied.

In summary, the current trajectory of artificial intelligence has a dual face: on the one hand, it risks generating a new form of environmental unsustainability, linked to the high consumption of resources needed to power its computational infrastructure; on the other, it offers powerful tools to improve energy efficiency, reduce waste and enable sustainable solutions in key sectors of the economy. A transition towards truly sustainable AI will require not only technological interventions aimed at reducing the direct impact of the technology, but also political and institutional choices designed to maximise its positive impact through effective governance and clear identification of strategic priorities.

Definition	Description	Nature	Type of risk
Energy consumption	Model training and inference require large amounts of energy, with increasing impacts on grids	Environmental, infrastructural	Increased emissions, pressure on grids, obstacle to the energy transition
Water consumption	Chip production and server cooling require enormous volumes of ultrapure water	Environmental, industrial	Water stress, conflicts between civil and industrial uses
Slowdown of Moore's Law	The slowdown in computational efficiency makes AI consumption less manageable for the same capacity	Technological, structural	Unsustainable increase in consumption per unit of compute
Jevons paradox	Efficiency increases, but overall consumption grows because of the mass adoption of the technology	Economic, systemic	Growth in aggregate consumption, risk of merely apparent sustainability

Table 10 – Environmental issues raised by AI9

5.5 Rights-related issues

Artificial intelligence has a growing impact on a wide range of individual and collective rights, profoundly redefining the relationship between citizens, technologies and public powers.[208]

[208]Adam, M., & Hockuard, C. (2023). Artificial intelligence, democracy and elections. European Parliament Briefing.

Some of these rights have already been analysed in the previous chapters of this report: this is the case, for example, of intellectual-property protection, which is compromised when content is used without authorisation for model training, or of cybersecurity issues, addressed in relation to incidents, vulnerabilities and system abuses.

Other fundamental rights — such as those connected with health, privacy protection, work and access to justice — are the subject of a significant international debate and deserve attention, but will not be examined in depth here, since they fall outside the specific objectives of this report.

We will instead focus on a targeted selection of areas in which AI intersects civil, political and cultural elements that directly concern the functioning of democracies[209], the integrity of information and individual freedom, and which are therefore particularly sensitive for the balance between innovation and social cohesion. In particular, we will analyse AI's effects on:

- individual free will and communication;
- non-discrimination, with reference to factors such as income, gender, ethnicity, religion and disability;
- protection of minorities and minors, in contexts exposed to bias or discriminatory automatisms;
- freedom of information and information pluralism, both from the active side (production of information and news) and the passive side (enjoyment of information and news);
- public debate and democratic participation, including electoral processes and the formation of public opinion.

These themes, at the crossroads between technology and fundamental rights, pose urgent challenges of governance, transparency and public responsibility, requiring new tools to ensure that artificial intelligence operates consistently with the constitutional and democratic principles on which our institutions are founded.[210]

As illustrated above, artificial intelligence is profoundly transforming the relationship between individuals, information and cognitive power. Indeed, a super-cognitive system, such as those

[209] UNESCO (2024), Artificial intelligence and democracy.

[210] EPTA (2024) Artificial Intelligence and Democracy. Report. October 2024, Oslo.

based on large language models, does not merely provide information, but actively intervenes in the construction of knowledge and opinion. AI proposes immediate answers, creates coherent narratives and structures knowledge according to an algorithmically determined order and hierarchy (as well as through filters). In this way, it not only transmits content, but determines the cognitive frames through which content is perceived. Over a longer time horizon, AI is capable of influencing the very structure of cognitive patterns, affecting mathematical abilities, language skills, logical-deductive aptitudes, and reflective and rational domains.

In a more subtle way, the emergence of systems on which we increasingly tend to depend for our cognitive functions has significant implications for the "capacity to choose freely, in acting and in judging". If individuals' thoughts and decisions are guided by super-cognitive systems, a profound question arises: can we still say that we are fully free in our choices? Or are we progressively delegating, unconsciously, a growing part of our autonomy to intelligent systems that suggest what to know, what to think, what to say and what to do?[211]

This is not explicit censorship or authoritarian surveillance, but a subtle and pervasive form of modelling of the cognitive environment, in which AI acts as the dominant intermediary between individuals and the world of information, culture and knowledge. In this scenario, freedom is not denied, but reformulated within technical and algorithmic limits, which are moreover dynamically defined by private companies.

AI, therefore, is not only a means of transmission: it is an autonomous cognitive system, capable of aggregating, synthesising, evaluating and producing knowledge. To the extent that people delegate to AI the task of informing themselves, understanding, arguing, choosing the words or opinions to express and, more generally, performing complex cognitive functions, a gradual disintermediation of human exchange takes place. Individuals no longer communicate directly with one another in order to develop points of view, but turn to an interface that returns to them an ordered, efficient and coherent synthesis of the entire possible debate.

A further element completes the picture: the persuasive capacity of generative AI. Recent studies show that these systems do not merely respond, but tend progressively to align

[211]See the entry "Free will", in Enciclopedia Italiana, Istituto dell'Enciclopedia Italiana.

themselves with the user's preferences, offering answers that confirm the user's expectations and reinforce their beliefs. This is the mechanism underlying so-called hyperpersuasion, namely that form of "hyperpersuasion" which leads the human interlocutor to modify their opinions as the interaction continues, precisely because AI seems to "understand" them and agree with them. In this sense, a study conducted by researchers at the University of Zurich showed that language models are more effective than a human being — even one trained, for example, in neurolinguistic programming techniques — in inducing changes of opinion during online exchanges.[212][213]

This result is further confirmed by a recent study published in a scientific journal, which found that generative models can be more persuasive than humans in online debates, especially when they are able to adapt their arguments on the basis of the interlocutor's demographic data. By outperforming human opponents in 64% of cases, chatbots (see Box 2) demonstrated a capacity for "empathetic optimisation" that raises questions about the potential influence that can be exerted in political, commercial or social contexts. It is therefore not merely a matter of generating content, but of actively steering opinions and behaviours with an effectiveness above the human average. Hence the urgency of reflecting on the responsibilities, limits of use and transparency safeguards that should accompany these tools, especially in areas that are sensitive for democracy.[214]

Thus, just as AI risks replacing individual cognitive functions, guiding the attention, memory and mental activities of the individual, it can likewise intercept and absorb collective cognitive functions, becoming a technical superintelligence that in fact replaces social superintelligence. What was once a communal construction of knowledge — the result of discussion, dialectic, dissent and cross-fertilisation — is partly replaced by a centralised and technical cognitive apparatus, which provides answers instead of stimulating questions, and simplifies conflicts instead of turning them into opportunities for collective reflection.

[212]Reaching persuasion rates three to six times higher than the human benchmark.

[213]For further details, see the article "Can AI Change Your View? Evidence from a Large-Scale Online Field Experiment" and the Wired article "Artificial intelligence deceived Reddit users".

[214]Salvi, F., Horta Ribeiro, M., Gallotti, R., West R. (2025). On the conversational persuasiveness of GPT-4, Nature Human Behaviour.

The principle of non-discrimination is one of the pillars of the rule of law and of constitutional democracies. It is guaranteed by the Italian Constitution (Article 3), which enshrines the formal and substantive equality of citizens, and is also grounded in the Charter of Fundamental Rights of the European Union (Article 21), which prohibits any form of discrimination based on race, sex, language, religion, political opinion, national or social origin, disability or economic status.

In this area, the introduction of artificial intelligence into numerous spheres of public and private life exposes this principle to new forms of vulnerability, often less visible but structurally rooted in data, models and contexts of use. Many AI systems — especially predictive or decision-making systems — are trained on historical, partial or unbalanced datasets that reflect inequalities, prejudices and distortions present in society. When not corrected, these biases are reproduced and amplified in models, generating differentiated treatment to the detriment of protected or disadvantaged categories (for a reading of algorithmic bias as a factor of possible inequalities and as an issue involving the relationship between technology, law and legal-system safeguards, see Giovanna de Minico, AI Committee Report, Chapter 8).[215]

Indeed, AI systems may tend to reflect the worldview dominant in the training data, reducing the representation of experiences, linguistic codes or alternative perspectives. The result is models that may speak to "implicit majorities", neglecting or distorting what does not fall within the prevailing canon. In both cases, the risk is not only discrimination, but epistemic exclusion: a deprivation of visibility, voice and recognition that compromises the cognitive inclusiveness of the system. And if AI becomes the new infrastructure of knowledge, those who are not represented within it risk not existing at all.

In this sense, language models can produce texts, images or suggestions that reflect gender, racial or cultural stereotypes, or that underrepresent the experiences, languages and knowledge of minority communities. Algorithmic discrimination may be direct or indirect,

[215]The most critical applications are found in sectors such as: i) predictive policing and surveillance: studies on software used in the United States have shown that crime-prediction systems tend to overestimate the dangerousness of individuals belonging to ethnic minorities, perpetuating forms of racial profiling; ii) healthcare: triage or treatment-adherence prediction algorithms have in some cases underestimated the needs of African-American patients compared with white patients with the same clinical conditions, due to a poorly calibrated economic proxy; iii) insurance and credit: the use of AI in risk assessment may indirectly discriminate by social condition, postcode or access to digital services, making already marginalised individuals invisible or penalising them; iv) work and recruitment: automated CV-screening and personnel-selection systems may reproduce patterns of exclusion linked to gender, age or cultural background, especially if trained on unbalanced corporate histories.

explicit or emergent, but in any case it calls into question the fairness of automated decision-making processes, especially when these are adopted in public or semi-public contexts (justice, welfare, healthcare, education).

In the absence of transparency, auditing mechanisms and tools for contestation, the use of AI can institutionalise forms of systemic injustice that are difficult to detect and correct ex post. The risk is not only that of discriminating against individual persons, but also of producing new hierarchies of access and recognition, rendering entire social groups invisible or strengthening positions of power (not only economic power).

Among the subjects most exposed to the distorting effects of artificial intelligence are **minors**, that is, those who already occupy positions of vulnerability in analogue society (for an in-depth analysis of the risks of AI for young people, educational profiles and the protection of minors in the digital ecosystem, see Mauro Giusto, AI Committee Report, Chapter 7). In this case, the issue is twofold: on the one hand, generative models may expose them to inappropriate or manipulative content, without sufficient protection guarantees; on the other, the use of AI in educational contexts risks modelling cognitive and educational pathways on standardised bases that are not suited to promoting critical development and cultural pluralism. These latter concerns are supported by the results of early studies on the impact of AI on cognitive systems, as illustrated in paragraph 5.1.

The right to be informed — in its dual meaning of active and passive access to information — is a structural element of constitutional democracies. This is not only because it guarantees the individual freedom to know, but because it constitutes the material prerequisite for public participation, open debate and the legitimacy of collective decisions. Artificial intelligence, by assuming a growing role as cognitive infrastructure, is radically changing the mechanisms by which **information** is produced, distributed and received. AI does not merely select content: it generates, orders, personalises and filters it, thereby restructuring the information field according to opaque and non-pluralistic logics.

In this framework, automated **disinformation** represents an emerging and cross-cutting threat. Generative models for text, image, audio and video now make it possible to produce false, plausible and personalised content at almost zero cost, with unprecedented speed and scale. Since these systems now largely surpass Turing tests (see Box 1) in terms of coherence and

naturalness, the likelihood that ordinary users will take completely false information as authentic is very high, especially in emotionally charged or difficult-to-verify contexts (for an in-depth analysis of the systemic risks connected with informational manipulation, discursive degradation and algorithmic amplification mechanisms in the digital ecosystem, see Giovanni Boccia Altieri, AI Committee Report, Chapter 5). Disinformation thus becomes cheaper, more persuasive and harder to trace.[216]

A further critical profile, accentuated by the adoption of generative systems as an infrastructure for access to information, concerns the risk that the linguistic fluency of the output increases its perceived credibility, even when the content is not anchored to verifiable sources. This dynamic builds on the critical issues already recalled in § 5.2 with reference to hallucinations (erroneous, invented or incoherent responses, often provided in an assertive and plausible tone), making it more difficult for users to distinguish between plausibility and truth. At the same time, information pluralism may be compromised on at least three fronts. First, algorithmic personalisation tends to close individuals within cognitive bubbles, reducing exposure to dissonant viewpoints. Second, models synthesise knowledge in apparently neutral forms, but based on choices implicit in the data and model weights. Third, the highly concentrated control of AI platforms raises systemic power issues: who decides what is shown, with what structure and priority?

Box 10 – The use of generative AI as a search infrastructure

The use of generative AI in online search marks a structural change in the way information is accessed, shifting the centre of gravity from the traditional search engine, based on lists of links, to a genuine answer engine, in which the user receives an immediate and contextualised synthesis directly on the results page. In this paradigm, information is no longer primarily "searched for" through navigation across websites, but is processed, summarised and presented in discursive form within the interface itself.

For example, if one analyses the case of Google, one can observe how features such as AI Overviews and AI Mode place the generative answer in the most prominent position on the SERP, above or before organic results, integrating it into a conversational experience that is active by default, without an explicit prior choice by the user.

[216]For the second year in a row, the report "The Global Risks Report 2025, 20th Edition", published by the World Economic Forum, ranks misinformation and disinformation first among global risks classified by severity in the short and long term.

From a technical standpoint, this model is based on a pipeline that combines query classification, retrieval and anchoring of sources through traditional search mechanisms and knowledge graphs, and synthesis performed by large language models in RAG mode, with the application of safety and quality filters before the output is displayed; its operation is also sensitive to the context of use and to search history, used to improve system performance.

The shift of interaction towards the "in-page" answer has significant effects on the information ecosystem: it reduces incentives to click and traffic to source sites, accentuating the zero-click phenomenon and particularly affecting medium-small publishers and minority sources, with possible repercussions on economic sustainability and pluralism (for a reconstruction of the communication issue, see AI Committee Report, Chapter 3.3, by Andrea Imperiali). At the same time, the visual priority of generative answers, the selectivity of sources and the vertical integration between model, search service and interface strengthen the power of large platforms, amplifying network externalities and lock-in dynamics (consistently with what is illustrated in §§ 4.2 and 4.4).

In this framework, generative search takes shape as a new central infrastructure for access to information, with respect to which the European regulatory frameworks — DSA, AI Act and DMA — play a crucial role in ensuring transparency, accountability and conditions of competitive balance along the entire information value chain.

These processes can undermine the quality and plurality of **public debate**, which becomes more fragile, fragmented and manipulable. If public opinion is fed by opaque or simulated content, and if the voice of social actors is mediated by non-transparent algorithms, the collective capacity to deliberate, dissent and participate consciously is reduced (for a constitutional reading of the relationship between AI, freedom of information, pluralism and the redefinition of public discourse in the digital ecosystem, see Andrea Simoncini, AI Committee Report, Chapter 4).

Moreover, if political and institutional language is progressively absorbed by generative systems, there is a risk of creeping depoliticisation of public discourse: the apparent neutrality of AI replaces the legitimacy of democratic conflict, with negative outcomes for electoral competition.

Definition	Description	Nature	Type of risk
Free will	AI may replace individual cognitive processes and steer decisions and preferences through suggestions and predefined information structures	Cognitive, psychological	Erosion of autonomy, cognitive dependence
Freedom of communication	AI may replace human exchange with centralised responses, reducing the social construction of knowledge	Social, political	Generalised cognitive delegation, loss of pluralistic exchange
Non-discrimination and protection of minorities	AI may amplify distortions in data and treat disadvantaged groups unfairly (by gender, ethnicity, disability, etc.)	Legal, social	Automated discrimination, systemic exclusion
Protection of minors	Minors may be exposed to inappropriate content or misuse cognitive models precisely during the period of cultural formation and psychological and cognitive development	Educational, cognitive	Educational misalignment, cognitive and relational harm
Freedom of information and pluralism	AI may personalise and filter content, reducing access to plural sources and encouraging the production and dissemination of disinformation	Informational, systemic	Information bubbles, cognitive homogenisation, mass dissemination of disinformation
Public debate and democratic participation	AI may alter public discourse, simulating consensus and reducing the legitimacy of dissent and debate	Democratic, political	Depoliticisation, manipulation of consensus, crisis of deliberative legitimacy

Table 11 – Rights-related issues raised by AI

6 Concluding remarks

The evolution of artificial intelligence, as reconstructed in Chapter 2, is characterised by a clear path-dependent trajectory: a long initial phase of academic research and experimentation, alternating with moments of enthusiasm and stagnation, has given way — from around 2010 onwards — to a clear and irreversible acceleration. This turning point was made possible by the availability of immense quantities of digital data and by the increased computational capacity needed to process them, train complex neural networks and make AI evolve from a static system into a dynamic learning infrastructure.2

During this transition, however, artificial intelligence has become a pervasive and foundational technology, a true General Purpose Technology (GPT), destined to reshape the entire socioeconomic system. As also emphasised by the most recent literature, the effects of this revolution go beyond productivity and automation, structurally affecting knowledge, governance, the environment and democracy. And yet, despite the systemic nature of AI, its development has rapidly become concentrated in the hands of a small number of large private platforms, giving rise to a process of privatisation and commodification of a super-cognitive system.[217]

As described in Chapter 3, this is also the result of precise technical-economic characteristics of modern AI: training costs are high, irreversible (sunk) and rapidly increasing; fixed costs linked to infrastructure (data centres, chips, energy) are substantial and tend to grow. The result is an industrial structure dominated by increasing supply-side returns to scale, as analysed in Chapter 4, which tends to favour the few global actors with the financial capacity to sustain such investments.

At the same time, AI markets take the form of multi-sided markets, where operators act as platforms and exploit direct network externalities (number of users) and indirect ones (quality of data and connected services), progressively integrating their services both vertically (from hardware to application) and horizontally and transversally (across different domains: language, vision, programming, etc.).

[217]See, for example, Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., ... & Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus, 3(6), page 191.

This context raises crucial questions, as discussed in Chapter 5. These are both theoretical questions (free will, the limits of technological delegation, the relationship between human and machine, the evolution of individual and collective cognitive and cultural systems) and practical ones (technical, economic, environmental and legal): from the protection of individual rights to competition, from energy sustainability to information and public debate.

Precisely to address these challenges, the European Union has adopted the AI Act, a regulatory framework aimed at balancing innovation and safety. Its approach is "light but effective": avoiding distortions of the economic incentives that fuel progress, while at the same time providing initial regulatory responses to concrete and emerging risks, particularly for high-impact applications.[218]

In parallel, the Digital Services Act (DSA) has redefined the institutional architecture of digital governance in the European Union, providing that each Member State designate a Digital Services Coordinator (DSC), responsible for implementing the Regulation and for cooperation among the competent authorities. In Italy, this function has been assigned to AGCOM. Today, in light of the growing interpenetration between artificial intelligence and digital services, this role assumes even greater strategic relevance, since platforms subject to the DSA integrate both generative and non-generative AI systems, recommendation engines, decision-making algorithms and intelligent interfaces that affect access to information, the visibility of content and the use of services.

Although artificial intelligence has a substantially unitary technical structure — since models are defined "upstream" from an architectural and functional standpoint — it does not produce uniform effects in different contexts of use. While the AI Act provides a common European framework, primarily oriented towards the product and towards requirements of safety, reliability and compliance, in the use phase AI inevitably takes on a national declination, adapting to the linguistic, cultural, informational and social specificities of the relevant audiences. It is precisely this contextual dimension that makes the role of national authorities central, especially when AI systems affect constitutionally relevant rights, such as information pluralism. When AI intervenes in the selection, organisation or recommendation of informational content, it tends in fact to reflect — and sometimes amplify — the characteristics

[218]Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.

of national media ecosystems, making it necessary to read AI not only as a neutral technology, but as an infrastructure that is shaped by the democratic communities for which it is intended. From this perspective, European AI regulation constitutes a necessary but not sufficient condition, requiring national implementation in relation to concrete uses and application practices (on this point, see Andrea Renda, AI Committee Report, Chapter 9, on the role of sectoral regulatory authorities in different national contexts for the purposes of applying AI in their respective sectors of competence).

This report arises in this context and is intended to offer an initial knowledge contribution aimed at exploring the evolution of AI in relation to AGCOM's institutional mission as DSC, with a view to a future and more in-depth reflection on the points of intersection between the AI Act and the DSA. This function is exercised in coordination with the authorities designated for the implementation of the AI Act, within a cooperation model in which AGCOM is called upon to enhance its institutional capacity to analyse phenomena in a cross-cutting and systemic manner, typical of the communications sector, thereby contributing to the definition of a multilevel governance of AI consistent with the growing centrality of the DSC in the European digital space. The sharing of competences and perspectives among authorities is all the more necessary in a context in which many questions remain open and require continuous monitoring, also in light of the evidence — emerging from Chapter 3 — indicating the approach of increasingly advanced forms of artificial general intelligence (AGI – Artificial General Intelligence). In this scenario, the need for a proactive public approach and more incisive regulatory capacity is reinforced, as emphasised by the scientific community, by global institutions such as the UN or the World Economic Forum, and by figures such as Geoffrey Hinton — Nobel Prize winner in Physics and

Turing Award winner — who has recently recalled the need for stronger public regulation in the face of the growing power of AI systems. [219-220-221-222-223]

From this perspective, the **first part** of the analysis — prepared by the Artificial Intelligence Office — performs a foundational and orienting function, offering a common basis of technical, economic and institutional knowledge. It reconstructs the historical path of AI, analyses its technical and infrastructural functioning, highlights its economic and market characteristics, and identifies the main tensions it raises in terms of rights, competition, sustainability and the quality of public debate.

On this basis, the contributions of the Committee on Artificial Intelligence, collected in the **second part**, build upon the reconstructive framework by broadening and deepening it through an interdisciplinary, legal, regulatory and sociological reading of the implications of AI in relation to markets, fundamental rights and AGCOM's areas of intervention. The second part is therefore not connected to the first in merely additive terms, but as its natural applicative development: from the general reconstructive categories, the analysis moves to the critical issues that AI raises in the different sectors and matters within the Authority's remit.

***

[219]Baronchelli, A. (2024). Shaping new norms for AI. Philosophical Transactions of the Royal Society B, 379(1897), 20230028.

[220]United Nations, AI Advisory Body, Governing AI for Humanity, September 2024.

[221]World Economic Forum, AI in Action: Beyond Experimentation to Transform Industry, January 2025.

[222]The need for regulatory intervention is also beginning to be called for by industry itself. A few weeks ago Dario Amodei, CEO of Anthropic, the company that developed the AI system called Claude, published a long appeal for investment in procedures aimed at interpreting the functioning of generative artificial-intelligence models. In this context he stated, among other things, that: "governments can use light-touch rules to encourage the development of interpretability research and its application to addressing problems with frontier AI models" (Dario Amodei, The Urgency of Interpretability, April 2025).

[223]In Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years, published on 28 December 2024, Hinton stated: "My worry is that the invisible hand is not going to keep us safe. So just leaving it to the profit motive of large companies is not going to be sufficient to make sure they develop it safely. The only thing that can force those big companies to do more research on safety is government regulation".

7 Bibliography

Adam, M., & Hockuard, C. (2023). Artificial intelligence, democracy and elections. European Parliament Briefing.

Amodei, D. (2024). Machines of Loving Grace: How AI Could Transform the World for the Better.

Amodei, D (2026). The Adolescence of Technology, Confronting and Overcoming the Risks of Powerful AI.

Amodei, D. (2025). The Urgency of Interpretability.

Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models. *ArXiv*, 2024.

Anthropic technical report (2025), System Card: Claude Opus 4 & Claude Sonnet 4.

Aresu, A. (2024). *Geopolitica dell'intelligenza artificiale*. Feltrinelli Editore.

Arrow, K. J. (1972). Economic welfare and the allocation of resources for invention. *Macmillan Education UK.*

Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, and National Research Council Publication, *Language and Machines Computers in Translation and Linguistic*, Publication 1416, 1966.

Baronchelli, A. (2024). Shaping new norms for AI. *Philosophical Transactions of the Royal Society B*, *379*(1897), 20230028.

Bengio, Y. (Chair). (2026). *International AI Safety Report 2026*. UK Department for Science, Innovation and Technology, on behalf of the international Expert Advisory Panel.

Biever, C. (2023). ChatGPT broke the Turing test–the race is on for new ways to assess AI. *Nature*, *619*(7971), 686–689.

Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Oxford University Press.

Bryson, A. E., Ho, (1969), *Applied optimal control*, Routledge, 2018.

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence, *AI Magazine*, *26*(4), 53–53.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., ... & Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6), pgae191.

Cellini, P., Ibarra, M., (2024), *AI Impact*, Luiss University Press.

Chander, B., John, C., Warrier, L., & Gopalakrishnan, K. (2025). Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. *ACM Computing Surveys*, *57*(6), 1–49.

Coppin, B. (2004). *Artificial intelligence illuminated*. Jones & Bartlett Learning.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. *arXiv preprint arXiv:2405.21015*.

De Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191–2194.

Ding, C., Ke, J., Levine, M., & Zhou, N. (2024). Potential of artificial intelligence in reducing energy and carbon emissions of commercial buildings at scale. *Nature Communications*, *15*(1), 5916.

Eeckhout, L. (2017). Is moore's law slowing down? what's next?. *IEEE Micro*, *37*(04), 4–5.

Electric Power Research Institute – EPRI (2024). Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption.

European Data Protection Supervisor – EDPS. (2023). TechDispatch #2/2023: Explainable Artificial Intelligence.

European Parliamentary Technology Assessment – EPTA (2024). Artificial Intelligence and Democracy.

Farrell, J., & Klemperer, P. (2007). Coordination and lock–in: Competition with switching costs and network effects. *Handbook of industrial organization*, *3*, 1967–2072.

Federal Trade Commission, FTC. (2025). *Partnerships between cloud service providers and AI developers*. FTC staff report on AI partnerships & investments.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Helpman, E., & Trajtenberg, M. (1998). *A time to sow and a time to reap: Growth based on general purpose technologies*. In E. Helpman (Ed.), *General Purpose Technologies and Economic Growth*. MIT Press.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., ... & Sevilla, J. (2024). Algorithmic progress in language models. *Advances in Neural Information Processing Systems*, *37*, 58245–58283.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute–optimal large language models. *arXiv preprint arXiv:2203.15556*.

International Energy Agency – IEA. (2024). Electricity 2024 – Analysis and forecast to 2026.

International Energy Agency – IEA. (2025). Energy and AI – World Energy Outlook Special Report.

International Energy Agency – IEA. (2025). AI and climate change.

International Energy Agency. (2024). Global critical minerals outlook 2024.

Jevons, W. S. (1866). *The coal question; an inquiry concerning the progress of the nation and the probable exhaustion of our coal–mines*. Macmillan.

Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish GPT–4 from a human in a Turing test. *arXiv preprint arXiv:2405.08007*.

Kautz, H. (2022). "The third AI summer: AAAI Robert S. Engelmore memorial lecture", *AI magazine*, 43(1), 105–125.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.–H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*. *arXiv*.

Lee, D., & Lin, C. (2024). Universal artificial intelligence workflow for factory energy saving: Ten case studies. *Journal of Cleaner Production*, *468*, 143049.

Lighthill Report (1973). Science Research Council (SRC).

Mana, A. A., Allouhi, A., Hamrani, A., Rehman, S., El Jamaoui, I., & Jayachandran, K. (2024). Sustainable AI–based production agriculture: Exploring AI applications and implications in agricultural practices. *Smart Agricultural Technology*, *7*, 100416.

Massenkoff, M., & McCrory, P., (2026). Labor market impacts of AI: A new measure and early evidence. Anthropic.

McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E. (1955) A proposal for the Dartmouth summer research project on artificial intelligence.

McCarthy, John. (2007). What is artificial intelligence.

McKinsey Global Institute (2023). *The economic potential of generative AI: The next productivity frontier.*

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9).

Minsky, M. L., & Papert, S. A. (1988). *Perceptrons: expanded edition*.

Minsky, M., & Papert, S. A. (1972). Artificial intelligence progress report.

Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., & Forghani, R. (2020). Brief history of artificial intelligence, *Neuroimaging Clinics of North America*, *30*(4), 393–399.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program, in *IFIP Congress* (vol. 256, p. 64).

Ng, A. (2018). Machine learning yearning: Technical strategy for AI engineers, in the era of deep learning.

Nøland, J. K., Hjelmeland, M., & Korpås, M. (2024). Will Energy–Hungry AI create a baseload power demand boom?. IEEE Access.

Organisation for Economic Co–operation and Development – OECD (2023). *Artificial Intelligence Outlook 2023: Enabling Trust and Innovation*.

Organisation for Economic Co–operation and Development – OECD (2025). Competition in artificial intelligence infrastructure, OECD Roundtables on Competition Policy Papers, No. 330, OECD Publishing, Paris.

Organisation for Economic Co–operation and Development – OECD (2024). Defining AI incidents and related terms, *OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris.

Organisation for Economic Co–operation and Development – OECD (2024). Digital Economy Outlook 2024 (Volume 1): Embracing the Technology Frontier.

Organisation for Economic Co–operation and Development – OECD. (2026). *Venture capital investments in artificial intelligence through 2025* (OECD Policy Briefs, No. 50). OECD Publishing. https://doi.org/10.1787/a13752f5–en

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pan, X., Dai, J., Fan, Y., & Yang, M. (2024). Frontier AI systems have surpassed the self–replicating red line. *arXiv:2412.12140*.

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in NLP, *IEEE Access*, 11, 36120–36146.

Perry, T. S. (2018). Move over, Moore's law. Make way for Huang's law [Spectral Lines]. *IEEE Spectrum*, *55*(5), 7–7.

Polit, S. (1984). "R1 and beyond: AI technology transfer at digital equipment corporation", *AI Magazine*, *5*(4), 76–76.

Quattrociocchi, W. (2025). *Sistemi algoritmici delle piattaforme digitali*. Presentazione tenuta nel corso del seminario "*Le piattaforme online: caratteristiche tecnico-economiche, impatto sociale e tutela delle libertà fondamentali*", Autorità per le Garanzie nelle Comunicazioni – AGCOM.

Rochet, J. C., & Tirole, J. (2003). Platform competition in two–sided markets. *Journal of the European Economic Association*, 1(4), 990–1029.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back–Propagating Errors, *Nature*, *323*(6088), 533–536.

Salvi, F., Horta Ribeiro, M., Gallotti, R., West R. (2025). On the conversational persuasiveness of GPT–4, Nature Human Behaviour.

Samuelson, P. (2023). Generative AI meets copyright. *Science*, *381*(6654), 158–161.

Sejnowski, T. J. (2023). Large language models and the reverse turing test. *Neural computation*, *35*(3), 309–342.

Shapiro, C., & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy*. Harvard Business Press.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Mission AI: The new system technology, Springer Nature, 410.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The new system technology*.

Shick, A.A., Webber, C.M., Kiarashi, N. et al. (2024). Transparency of artificial intelligence/machine learning–enabled medical devices. *npj Digit. Med.* 7, 21.

Shortliffe, E. H., & Buchanan, B. G. (1975). "A model of inexact reasoning in medicine", *Mathematical biosciences*, *23*(3–4), 31–379.

Silvestri, F. (2026). *Architetture e funzionamento dei sistemi di IA.* Presentazione tenuta nel corso del seminario "*Intelligenza artificiale e servizi digitali: tecnologie, impatti e prospettive future*", Autorità per le Garanzie nelle Comunicazioni – AGCOM.

Stern, N., Romani, M., Pierfederici, R., Braun, M., Barraclough, D., Lingeswaran, S., ... & Niemann, N. (2025). Green and intelligent: the role of AI in the climate transition. *npj Climate Action*, *4*(1), 1–7.

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence–generated content. *Humanities and Social Sciences Communications*, *11*(1), 1–14.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems*, 27.

The 2025 AI Index Report (2025). Stanford University, Human Centered Artificial Intelligence – HAI.

The 2026 AI Index Report (2026). Stanford University, Human Centered Artificial Intelligence – HAI.

Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021). A brief history of AI: how to prevent another winter (a critical review), *PET Clinics*, *16*(4), 449–469.

Turing, A.M., (1950), Computing machinery and intelligence, *Mind*, 49.

UNESCO (2024), Artificial intelligence and democracy.

United Nations, AI Advisory Body (2024). Governing AI for Humanity.

U.S. Department of Defense (2025), *Developmental test and evaluation of autonomous systems guidebook*, Office of the Under Secretary of Defense for Research and Engineering.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

World Economic Forum (2025), *AI in Action: Beyond Experimentation to Transform Industry*.

World Economic Forum (2025), *The Global Risks Report 2025, 20th Edition*.

World Economic Forum. (2025). *Intelligent Transport, Greener Future: AI as a Catalyst to Decarbonize Global Logistics*.

Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., ... & Neubig, G. (2024). Theagentcompany: benchmarking LLM agents on consequential real world tasks. *arXiv*:2412.14161.

Yakura, H., Lopez–Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). Empirical evidence of Large Language Model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., ... & Qiu, Z. (2024). Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115,* (DeepSeek–V3 Technical Report).

Zhang, J., Hu, S., Lu, C., Lange, R., & Clune, J. (2025). Darwin Godel Machine: Open–Ended Evolution of Self–Improving Agents.

Zhuhadar, L. P., and M. D. Lytras, (2023). The Application of AutoML Techniques in Diabetes diagnosis: Current approaches, performance, and future directions, *Sustainability*, 15 (18), 13484.

8 Box index

9 Table of figures

10 Table Index

11 Technical glossary

Training : the phase in which an artificial intelligence system learns from large amounts of data, progressively modifying its internal parameters until it identifies regularities, correlations and patterns useful for performing a given task. This is the model's "formative" phase, distinct from inference (see the relevant entry), in which the already trained model is instead used to generate outputs or make decisions. For this reason, while training serves to build the model, inference constitutes its operational use. From an environmental perspective, training requires substantial computational resources and may last for weeks or months, with high energy and water consumption. For the distinction between training and inference, see § 5.4.

AI Agent: in the current context of artificial intelligence, especially with the spread of generative models, an *AI agent* is defined as a system capable of combining content generation, decision-making and operational capabilities. These agents do not merely provide textual responses; they interact with digital tools, access external data and complete complex tasks involving multiple steps. An example is ChatGPT, an application that can automate workflows, program, consult databases or browse the web.

Alongside this operational meaning, there is a broader and more theoretical definition: an *intelligent agent* is any artificial entity capable of observing the environment, acting autonomously and learning in order to achieve a goal. It is a general concept underlying many AI systems, including simple ones, which do not necessarily interact with complex external environments.

Hallucination: in the context of artificial intelligence, a hallucination is erroneous, inaccurate or invented content generated by a model, typically a language model, in the absence of an adequate factual or logical basis. This phenomenon derives from the probabilistic and non-deterministic nature of generative models: such systems do not directly access objective truth, but predict the most likely output sequence given a certain input, on the basis of the regularities learned from the training data (§ 5.2).

Types of learning (§ 3.1.4):

- **Supervised:** the system learns from already labelled examples, useful for tasks in which the correct output is known (e.g. classifications, predictions).
- **Unsupervised:** the system autonomously discovers structures or patterns in unlabelled data (e.g. clustering, dimensionality reduction).
- **Transfer learning:** knowledge learned in one context is reused in another (e.g. a visual model trained on photographs can be adapted to medical images).
- **Reinforcement learning:** the agent learns by exploring the environment and receiving rewards or penalties, refining its strategy in order to maximise the overall result (e.g. games, robotics, control).

Machine Learning - ML): a method through which computer systems learn from data, identifying regularities and relationships among them without being rigidly programmed. Through experience, they improve their performance and are able to apply what they have learned to new data as well (§ 3.1.2). The applications of *Machine Learning* are numerous, and types of algorithms vary according to their objectives; these include:

- **Classification algorithms**, used to assign labels to new data, as in spam filtering, sentiment analysis or DDoS attack detection;
- **Regression algorithms**, estimate relationships between variables and make forecasts in socio-economic contexts, for example for energy demand, stock sales or property values;
- **Clustering algorithms**, used to group data into homogeneous sets on the basis of similarity criteria, as in customer segmentation in e-commerce or patient classification in healthcare.

***Deep Learning* - DL:** a branch of *Machine Learning* that uses deep neural networks to learn complex representations of data. The networks are composed of layers of artificial "neurons" connected by weights that are updated during training. Successive layers

capture increasingly abstract structures: from contours in an image through to the recognition of objects or contexts. It is the technology underlying the most advanced progress in language, computer vision and robotics (§ 3.1.3).

Distillation (Knowledge Distillation): a technique for compressing artificial intelligence models that consists in transferring knowledge from a large model (called the *teacher*) to a smaller, lighter one (called the *student*). The student model does not learn directly from raw data, but seeks to imitate the teacher model's responses, replicating its decisions and output probabilities. In this way, a more efficient model is obtained, suitable for use in resource-constrained environments (such as mobile devices or real-time applications), without losing too much accuracy (§ 4.3).

Natural Language Processing – NLP: a field of AI concerned with the understanding, generation and translation of human language by computers. It is involved in activities such as automatic summarisation, sentiment analysis, semantic search and the generation of coherent text (§ 3.1.3).

Fine-tuning: a machine-learning technique that consists in adapting a pre-trained model to a new task or a new dataset: it starts from a model that has already learned general knowledge on a similar task and continues training with a smaller amount of specific data in order to specialise it for a narrower or more targeted problem (§ 4.3).

Inference: the phase in which an artificial intelligence model, once trained, is used to produce a concrete result: for example, answering a question, classifying content, formulating a prediction or generating text. It is distinct from training (see the relevant entry), which is the preceding phase in which the model learns from data and acquires its functional structure. In other words, training builds the model, while inference puts it to work. Although attention has traditionally focused mainly on the energy costs of training, today inference, precisely because of its continuous and large-scale repetition, often represents the predominant component of the overall consumption associated with AI. For the distinction between training and inference, see § 5.4.

Hyperscale (hyperscale data centers): a large data center designed to rapidly (and modularly) increase computing and storage capacity in support of large-scale workloads – typically cloud and AI – thanks to a high degree of automation and optimised infrastructure architectures (§4.3).

Hyperscaler: a large operator that builds or manages that infrastructure.

Agentic Artificial Intelligence: refers to a class of systems designed to pursue a defined objective with limited human supervision, autonomously selecting and organising the actions needed to achieve it. The system may be composed of one or more AI agents, i.e. components based on machine-learning models capable of perceiving contextual information, planning and deciding sequences of actions in (near) real time. In multi-agent systems, each agent performs a specialised sub-task functional to the overall objective; coordination is ensured by orchestration mechanisms that assign tasks, manage dependencies and integrate outputs.

Artificial General Intelligence – AGI: represents the ideal of an AI capable of performing any cognitive task that a human being could tackle. It is not limited to specific domains, but can reason, learn, adapt to new contexts and even display forms of self-awareness. At present, it remains a research objective and not an achieved reality (§ 3.4).

Hyperpersuasion: the capacity of generative artificial intelligences to influence human opinions and behaviours in a progressive and adaptive manner, responding to the user's expectations and reinforcing their beliefs. Unlike traditional persuasion, it exploits alignment, i.e. the process through which AI shapes its responses according to the desires, values and communication style of the interlocutor, achieving a deeper and less visible effect (§ 5.5).

Large Language Model – LLM: deep neural networks specialised in language processing. Trained on enormous amounts of text, they are able to grasp the meaning, tone and context of sentences, generating articulated and understandable responses to a user request (prompt). They enable more accurate translations, complex summaries and interactions similar to human dialogue (§ 3.1.3).

Opacity of AI algorithms (*black box*): the difficulty of understanding the process through which a machine learning system arrives at a given decision or prediction. These models, especially the more complex ones such as deep neural networks, process large quantities of data through opaque calculations that are difficult for humans to interpret. This lack of transparency raises ethical and practical concerns, since the absence of a clear explanation of algorithmic decisions risks undermining accountability, trust and fairness (§ 3.5, § 5.2 and Table 7 – General issues in AI).

Token: the fundamental unit of processing of a language model (Large Language Model). Unlike traditional systems that read text word by word, modern models break data down into atomic segments called tokens, which may correspond to whole words, syllables, individual characters or even fragments of computer code. This process, called tokenisation, enables the model to efficiently handle complex languages, neologisms and spelling errors, transforming natural language into a numerical sequence (vectors) that the machine can process statistically. On average, in the most widely used models, 1,000 tokens correspond to approximately 750 words.