Pedagogika, 2025, 75(4), 309–340
http://pages.pedf.cuni.cz/pedagogika/

A Brief History of Definitions of Artificial Intelligence: From the Great Bombe to Black Boxes

Tomáš Zemčík

Abstract: *The study explores and highlights the direct relationship between contemporary knowledge, paradigms, aims, and public expectations in the field of artificial intelligence (AI) and its definitions. For the purpose of this research, the division of the stages of the development of AI was used for analogies to the seasons; spring to winter of AI. The history of AI is covered here only in the range necessary to point out this relationship with the possibilities of deriving the resulting definitions. Examples of period-typical AI definitions are given for each period. This historical excursion is then used as a background for thinking about the form (not the content) of the definition of AI that is appropriate to the current state of the field and its paradigm, focus, and use. The current discussion on the shape of the definition of AI within the framework of EU legislation is outlined. The form of a suitable definition of AI for the present is examined from the perspective of interested parties, such as multinational entities, business organisations, and other stakeholders, and is compared with some already valid definitions of these entities. A paradox in the definitions of AI, which are always too "narrow and broad at the same time" from a certain point of view, is pointed out. Finally, the possibility of deploying a fractal definition with a fixed rational-moral core but changing content with respect to the levels at which it is applied is explored within a conceptual ideation. This operational fractal definition could, in principle, resolve the ever-present "broadness-narrowness" paradox.*

Keywords: *artificial intelligence; definitions; black box algorithms; history; regulation and ethics; cognitive bias*

1. Introduction

The term *Artificial Intelligence* (hereinafter referred to as AI) has become part of the vocabulary of the vast majority of us. At first glance, this is a clearly understandable term. The opposite is true. We can use it as an example to explain the so-called *Dunning-Kruger effect* (Johnson et al., 2013); mental automatism, which causes the general public to have the impression that they understand the topic more, and on the contrary, experts underestimate their own competence in the

DOI: 10.14712/23362189.2025.5009

field. And this is exactly the case with the definition of AI. Below, we will connect the definitions of AI with the contemporary and field context. In the text, the reader will find the resources to be able to answer the question: "Why is it important for society as a whole to know how AI can be defined?"

From this perspective, AI is a fundamentally illusory term: it refers not to actual intelligence, but to an engineered imitation – a surface-level reproduction of intelligent behaviour without the underlying cognitive substance, "*artificial imitation of intelligence*". Below, I will use the term "AI" in accordance with this thesis. For the purposes of this article, I will clarify the necessary context of this claim. The notion of AI as "Artificial Imitation of Intelligence" highlights the fact that AI is not genuine intelligence, but rather a metaphorical construct – an imitation or illusion of intelligence, where algorithmic systems replicate intelligent behaviour on the basis of functional relationships and observable output. It does not resemble human intelligence in any meaningful sense – it is basically just a sophisticated application of mathematical statistics, formal logic, and probabilistic models (Mikolov, Joulin, & Baroni, 2018). Moreover, the commonly-made distinction between narrow AI – systems designed for specific tasks – and general AI (often referred to as Artificial General Intelligence or AGI), which hypothetically could adapt to unfamiliar environments, make decisions under conditions of uncertainty, and operate across diverse domains, can be conceptually confusing and misleading. It is possible to define this difference, for example, according to ability, domain, task, or output or according to capability, which points to the ability, accuracy, or efficiency of the system – their level, quality, or distribution determines the generality of the system (Gutierrez et al., 2023). Abandoning the rigid distinction between narrow and general AI can enable a more liberated and nuanced approach to defining the concept of artificial intelligence.

For readers trying to make sense of the various definitions of artificial intelligences without extensive prior knowledge, it is important to highlight the often-unstated distinction between artificial and synthetic intelligence. Both human and animal intelligence are always completely different from automated advanced mathematical statistics represented by two-valued transistor silicon circuits calculating within two-valued formal logic. Intelligence is a part of existence, the realisation of personal and general consciousness and human (or animal) being experienced personally in the network of symbolic relations of the particular world or concept of reality. Yes, AI can address tasks similar to those performed by natural intelligence, but it does so in a fundamentally different way – often yielding results that diverge significantly from human cognition. Part of functional AI solutions in many cases must be to be accepted by people. This means that its outputs appear "as if presented by a being gifted with a feeling for

human consciousness" – as if its solution belongs to the human world, independent of the hyper-efficiency of the computational speed with which the machine arrived at the result. This can be seen, for example, with the popular ChatGPT from OpenAI. This AI does not generate outputs that are true but instead illusionary credible ones. Those who confuse *the illusion of trustworthiness with truthfulness* can be unpleasantly surprised.

And so the initial intention to conceive of AI as an artificial imitation of intelligence, regardless of the specific definition, means to disenchant the mechanisms of AI closed in a black box – to clarify the context of illusoriness often mistaken for reality in the sense of truthfulness. In the 21st century, in all layers of society and socio-cultural relations, both lay and professional, we need always to be aware that the outputs generated through AI mechanisms are always the solutions presented by the algorithms. These outputs come from systems that lack any form of human empathy, ethical awareness, or moral consciousness. Let us assume that these are not persons, but artificial technological prostheses (of mind?), *without the possibility of compromising one's person or corrupting a fragile soul*. These are technological products that do not consciously experience the social rules of the socio-cultural norms of the world into which they were thrown as beings. And they are also not subjects with an essential desire for goodness-beauty-utility!

2. Methods

In the academic literature, numerous definitions of AI can be found. However, a generally accepted definition remains elusive. As Tegmark (2018) notes, even the experts at a symposium convened by the Nobel Foundation were unable to reach consensus on a shared definition of intelligence – and later of artificial intelligence. And so, we can perhaps only agree on *the fact that there are different definitions*, *created for different purposes and with different goals, describing different aspects of the issue*. Each of these definitions reflects the specific goals, contexts, and interests of the individual authors or institutions that formulated them – shaped by their disciplinary orientation, the historical moment, and the intended applications. As a result, the concept of AI takes on slightly different meanings across definitions. The underlying assumptions, logical structures, and intended functions embedded in these definitions also vary accordingly.

The following examples help illustrate this point: the physicist and AI researcher Max Tegmark (2018) defines intelligence broadly as "the ability to achieve complex goals", in order to include as many future entities as possible – including artificial ones. Legislative and politico-economic definitions, on the other hand, tend to be narrowly focused on the specific domains to which the given regulation, policy, or funding applies (European Commission, 2018).

As cited in Barták (2017), one of the founders of the field of AI, Marvin Minsky, described AI as the science of creating machines or systems that use procedures which, if performed by a human, would require intelligence. According to Bostrom (2016), John McCarthy – who coined the term *AI* – pointed out that once something begins to work reliably, we tend to stop referring to it as AI. McCarthy thus highlighted a persistent tendency to perceive AI not as a present reality, but as a moving target always situated in the future – as part of futurology rather than the status quo. According to Dr. Andrew W. Moore, Director at Google Cloud AI, "AI is the science and research that enables computers to behave in ways that until recently we thought only human intelligence could" (Toosi et al., 2021). Once a particular technology is widely adopted and integrated into everyday use, it is often no longer labelled as AI, but instead given a specific name – such as a computer game or a text prediction tool. This phenomenon, referred to by Barták (2017) as the *AI effect*, highlights the tendency to perceive AI as something perpetually novel or futuristic, rather than part of the present technological landscape. Numerous other examples could be cited to illustrate how definitions are conditioned by the specific aims, contexts, and interests of those who formulate them. Arguably, such variability is not a flaw, but an intrinsic feature of the act of defining itself – shaped by the intentions and purposes it seeks to fulfil.

2.1 Defining AI in the Context of 2025

It seems reasonable to abandon the pursuit of a universal definition, or a universalist definitional paradigm grounded in the correspondence theory of truth – where a definition is expected to be the most accurate possible reflection of the entity that is being defined. Instead, a new paradigm should be anchored in a more appropriate framework of multi-layered, probabilistic definitions defined in a way that is perhaps akin to a definition with fractal dimensionality. In such an approach, different aspects of the phenomenon that is being defined are captured through distinct definitions, each addressing a particular layer of meaning and generating different implications and operational logics, while still sharing a logically coherent core.

The classical Aristotelian-Socratic model of definition – composed of *definiendum* and *definiens* – appears insufficient for the digital age of the early 21st century. Aristotle famously wrote: "A 'definition' is a phrase signifying a thing's essence. It is rendered in the form either of a phrase in lieu of a term, or of a phrase in lieu of another phrase; for it is sometimes possible to define the meaning of a phrase as well. People whose rendering consists of a term only, try it as they may, clearly do not render the definition of the thing in question, because a definition is always a phrase of a certain kind." (Topics I 5, 101b38–102a1). Given

the complexity of the AI phenomenon, it is clear that we must seek alternative approaches to defining it.

This type of layered definition aligns with the operational definition used by AI Watch (Samoili et al., 2020), provided that a higher degree of granularity is introduced into how we conceptualise AI. Each of these "granules" or definitional layers carries its own internal logic and requires its own ethical considerations. At the same time, these layers are not isolated – they intersect, influence one another, and share a common core: a commitment to moral responsibility and factual coherence. In principle, we should abandon the pursuit of a single, exclusive, interdisciplinary baseline definition of AI that claims universal validity across disciplines, speculative futures, and unpredictable technological developments. Instead, what is needed is a set of context-sensitive operational definitions – each tailored to specific environments and strategically applicable in ways that support the intended (and hopefully beneficial) goals of those shaping the AI ecosystem.

On the basis of my many years of close study of this field, I am reasonably convinced that such a multi-level definitional framework is essential. It enables us to distinguish between different layers, aspects, and circumstances of the application of AI – many of which operate according to incommensurable principles and therefore demand incommensurable ethical, legal, and economic approaches.

3. Predecessors of AI

The precursors of AI are largely related to the human (animal) desire to expand one's capabilities in manipulating reality – both internal and external. Therefore, we can place the beginning of AI in the context of the "history of human prosthetics". It is a gradual development of techniques and technologies, in which man goes beyond his natural attributes, for example, in terms of animal strength or reserves of energy management. An example can be a stick, stone, spear, or fist wedge as arm prostheses. Or fire expanding the possibilities of metabolism, when we can get more energy from the same amount of food after roasting. Fire also provides the relative safety of the daytime world, thereby extending it. Writing can be seen as a prosthesis of memory and an extension of informational presence to future levels. Floridi (2014) highlights that AI essentially continues the historical trajectory of enhancing human faculties through technological prosthetics – now extending beyond the body to encompass the mind, consciousness, will, agency, and intelligence.

Homer, who wrote of mechanical "tripods" waiting upon the gods at dinner, is often cited as the cultural origin of the idea of autonomous mechanical helpers. Aristotle, in *Politics* (Book I, 1253b), already discusses the essential need for both living tools with a soul and lifeless tools without a soul, thereby providing philosophical justifica-

tion for slavery. This concept becomes deeply anachronistic in an era in which the functions of so-called soulful tools no longer need to be performed by humans, but can be carried out by machines driven by algorithms (Aristotle, 1905).

We can trace other early manifestations of this idea back to the Enlightenment. René Descartes was intrigued by the notion of a machine capable of thought, although he did not pursue its practical realisation. In contrast, Gottfried Wilhelm Leibniz not only recognised the immense potential of mechanical reasoning machines based on logical principles – for instance, in resolving disputes – but also firmly believed in their technological feasibility, grounded in what he saw as "mechanical" relationships. Leibniz and Blaise Pascal constructed mechanical arithmetical calculators. However, it would be absurd to claim that these simple mechanical interlocking parts are capable of thinking, and therefore the label "artificial mind or intelligence" was out of place – it would have been completely absurd and inappropriate. This changes in the concept of Etienne Bonnot and Abbé de Condillac, who used the metaphor of a statue into whose head we pour particles of knowledge. In this thought experiment, they asked: "When will a statue know enough to appear intelligent to an observer?"; it is a figure of thought similar to one used later by Alan Turing (Buchanan, 2005).

The term *robot* originates from Karel Čapek's 1920 play *R.U.R.* (*Rossum's Universal Robots*). The name "Rossum" is a deliberate play on the Czech word *rozum*, meaning "reason" or "intellect", reflecting the play's philosophical themes and highlighting the focus on artificial rational beings. However, L. Frank Baum had already introduced a mechanical man named Tiktok in 1907. Baum describes Tiktok as extremely quick-witted, capable of forming independent thoughts, and perfectly articulate – a mechanical man who thinks, speaks, acts, and performs all functions except living. Tiktok's description could easily serve as a naive, early definition of artificial intelligence. It is also necessary not to forget the cultural inspirations of Jules Verne, Isaac Asimov, Mary Shelley, or the Jewish story about the Golem. The illusory Mechanical Turk chess machine, operated in the 18th and 19th centuries, was also a fascinating event. However, it was not until "modern computers" that the first realisations of these ancient visions, fantasies, and dreams of AI were really made possible. This was made possible, among others, by inventions in the laboratories of Alan Turing in Manchester, Howard Aiken at Harvard, Bell Laboratories, IBM, and the Moore School in Pennsylvania. In those days, the metaphor of computers as gigantic brains was generally used, and the definitions responded to this metaphor – they were directly derived from it (Buchanan, 2005).

4. The Cultural Seasons of AI: Spring, Summer, Autumn, and Winter

Some authors (e.g. Haenlein & Kaplan, 2019; Schuchmann, 2019; Toosi et al., 2021) have described the evolution of AI using metaphorical cycles inspired by the seasons – for example, spring, summer, autumn, and winter — to capture recurring waves of enthusiasm, maturity, decline, and stagnation in the field. For the purposes of this work, the model of four seasons was chosen.

4.1 AI Spring: The birth of AI

Alan Turing, a British mathematician, logician, and cryptanalyst, made foundational contributions to what would later become modern computer science. In the late 1930s, his focus was on rigorously defined theoretical problems, in contrast to the symbolic and fictional portrayals of artificial thinking beings such as R.U.R., Tiktok, the Golem, or Frankenstein's monster. For the British government, he developed a decryption machine called the Bombe (Polish for "bomb"), goal of which was to break the Enigma code used by the German armed forces during World War II. The Bombe, which was about 2.1×1.98×0.61 metres in size and weighed about a ton, is generally considered to have bene the first working electromechanical computer. In 1950, Turing published his seminal paper "Computing Machinery and Intelligence" (Turing, 1950), in which he described how to create intelligent machines and especially how to test their intelligence; known as the Turing machine. The name AI was still not being used at the time. It was not created until six years later at an event known as the Dartmouth workshop (The Dartmouth Summer Research Project on Artificial Intelligence or DSRPAI). A young associate professor of mathematics, John McCarthy, who was then at Dartmouth College, decided to organise a group to clarify and develop ideas about thinking machines. He chose the name "artificial intelligence" for the new field. The goal of DSRPAI was to bring together researchers from different disciplines and create a new area of research focused on creating machines capable of simulating human intelligence. Here, those who were later considered to be the founders of the field of AI combined their research efforts (Haenlein & Kaplan, 2019).

Berglund et al. (2023) show that nowadays – and most probably in the years to follow – the widespread adoption of chatbots powered by large language models (LLMs), based on artificial neural networks, has rendered the traditional Turing test largely obsolete. Many publicly available chatbots can now pass it effortlessly. However, according to the so-called Artificial Intelligence Effect – the idea that "once it starts working, we stop calling it artificial intelligence" – these chatbots, once integrated as everyday tools for managing routine tasks, are no longer perceived as *that* AI of which the advent is still projected into the future.

Today, less superficial evaluations – such as situational awareness tests (awareness to perception, situations, adaptation, prediction), theory of mind tasks (social and emotional intelligence), or context-sensitive interaction assessments (context) – are increasingly being used instead of the Turing test. But even in these cases, algorithms are beginning to fulfil or approximate the criteria set by such measures (Berglund et al., 2023).

4.1.1 The Turing test and the first definition

Turing (1950) seeks answers to the fundamental questions: "What does thinking mean and can machines think?" The question of whether machines can think is very difficult to answer, so he replaces it with the more pragmatic question: "Can a computer communicate in such a way that it is indistinguishable from a human?" To find the answer to such a problem, a simple controllable criterion or test, known today as the Turing test, can be used. This test is based on what Turing called the Imitation Game, a thought experiment in which a machine attempts to imitate a human in written conversation well enough to be indistinguishable from one. Thus, in this test, the ability to think is equated with the ability to communicate at such a level that the AI gives the impression to the participants in the conversation that it is a thinking being.

Introducing the first attempts to define AI in the context of the Turing test

However, the field of AI is not only about thinking machines. It is largely concerned with understanding intelligence and thinking as the manipulation of symbols or characters at a general level. In Turing's time, thinking was conceived of quite mechanistically – as a form of formal symbol processing. This reductive view has since experienced several renaissances, particularly during times of optimism in symbolic AI, and remains at the centre of ongoing debates: is thinking merely a matter of logic and formal rules, or does it fundamentally transcend such structures? This tension is clearly illustrated in the contrast between the GOFAI (Good Old-Fashioned Artificial Intelligence) approach and critics of such simplifications, such as Hubert Dreyfus, whose phenomenological critique articulated in 1972 (developed and extended in his 1992 work) – inspired by Heidegger – challenges the very premise that cognition can be reduced to rule-based symbol manipulation (Dreyfus, 1992). We can see the "Turing's time approach" in Herb Simon's statement from 1944: "Any rational decision may be viewed as a conclusion reached from certain premises…. The behavior of a rational person can be controlled, therefore, if the value and factual premises upon which he bases his decisions are specified for him." (Buchanan, 2005, p. 54)

The Turing test can be seen as an early operational definition of machine intelligence, formulated before the very concept of AI had been formally established. The term *AI* itself would not be coined until six years later, in 1956, at the Dartmouth Conference organised by John McCarthy

and his colleagues – widely regarded as the symbolic birth of the field. However, it does not attempt to define AI in a systematic or theoretical way – it was simply too early for such a formulation in Turing's time. Instead, Turing focused on more practical aspects of the problem, aiming to demonstrate how a machine could imitate human intelligence through computation (e.g. in relation to the *Entscheidungsproblem*, the halting problem, or questions of undecidability). Turing (1950) was intellectually rigorous and open about the potential objections to his proposed test, and he engaged critically with many of them. He famously predicted: "I believe that in about fifty years' time it will be possible to programme computers … to play the imitation game so well that an average interrogator will not have more than a 70 percent chance of making the right identification after five minutes of questioning."

4.2 AI Summer and Winter: The golden age and rise and fall of AI

The Dartmouth workshop not only established the field and introduced the concept of artificial intelligence; it also initiated a period of almost twenty years during which there were significant achievements in the newly-established field. Turing's optimistic vision became generally shared not only in science, but also in the media and pop culture. This period, later recognised as the first AI boom, ended with the funding crises known as the first AI winter.

In 1966, the American ALPAC (Automatic Language Processing Advisory Committee) report was published, which was very sceptical about the expenses and expected benefits in the field of AI, and yet in 1970, Marvin Minsky gave an interview to Life magazine stating that a machine with the general intelligence of an average human could be developed within three to eight years. But three years later, the British mathematician James Lighthill published a report commissioned by the British Council for Scientific Research, in which he once again questioned the optimistic outlook of researchers in the field of artificial intelligence. Lighthill stated that machines would only ever reach the level of an "experienced amateur" and only in games such as chess. Common sense reasoning would always necessarily be beyond their ability. In response, the British government ended its support for AI research at all universities (except Edinburgh, Sussex, and Essex). As a result of the ALPAC and Lighthill reports, and following the British example, the US Congress began to sharply criticise the high level of spending on AI research with little social benefit and virtually cancelled its funding. Thus, general scepticism regarding the field began to spread throughout society and practically worldwide, which started the phase sometimes called the AI Winter (Haenlein & Kaplan, 2019).

4.2.1 Discussion of the development of AI in the 1950s and 1960s

AI has never ceased to be interdisciplinary and will not cease to be, which must

necessarily be reflected in the disparity of definitions and its intentions and interpretations. Even in the years when the field of AI was still being formed, it was already strongly interdisciplinary. It extracted and incorporated knowledge from disciplines such as cybernetics, biology, experimental philosophy, communication theory, game theory, mathematics, statistics, logic, philosophy, psychology, linguistics, and others. AI outgrew these fields and influenced them in turn, setting up a reciprocal relationship of knowledge growth in these fields that continues to this day, resulting in today's versions of chatbots based on NLP principles (Natural Language Processing), which help in research and the training of new scientists, such as Google AI, Scite, Sci space, or ChatGPT.

The first definitions of AI are structurally linked to an idea that turns out to be more of a metaphor than a model or analogy. Researchers believed that the brain is basically just an extremely powerful computer, and simulating all its functions is basically just a matter of sufficient computing power, uncovering logical functional relationships, and enough data. In the 1950s and 1960s, we could see how the paradigm of the computer as a giant brain asserted itself. Through these technologies, humanity began to solve many problems that up to that time could only be solved by beings gifted with intelligence. However, because of the complexity of the topic, the research relies more on heuristics.

Feigenbaum and Feldman's 1963 volume *Computers and Thought* was one of the first major anthologies to bring together foundational texts in the emerging field of artificial intelligence. Five years later, Marvin Minsky (1968) published *Semantic Information Processing*, a comprehensive collection summarising much of the significant work carried out in the first decade after 1950. He emphasised that the central achievement of this period was the development of heuristic methods to constrain the search space in problem solving – trial and error guided by rules of thumb – which preceded the emergence of effective machine learning techniques. This heuristic focus opened the door to addressing deeper problems of knowledge representation and helped to make it possible to move beyond the rigid formalism of earlier logic-based systems. Nevertheless, formal heuristics remain a traditional and still relevant approach within the field (ibid.).

Minsky's work in the field of knowledge representation in AI essentially defined the formal theory of AI and the definitions derived from it. Essentially, it is the study of how the beliefs, intentions, and judgments of an intelligent agent can be appropriately expressed within a logic suitable for "automated reasoning" (Buchanan, 2005).

In the 1960s, AI models were inspired by important psychological questions and experiments. Practical applications were essentially implemented on the basis of rule-based programming. It was an attempt to simulate human manipulation with long-term and short-term memory features, including errors in reasoning.

E.g. Feigenbaum's EPAM program, completed in 1959, investigated associative memory and forgetting in a program that replicated the behavior of subjects in psychological experiments (Feigenbaum & Feldman, 1963).

During the 1960s, critical limitations of the symbolic AI approach became increasingly apparent, particularly in areas such as machine translation and understanding natural language. These shortcomings strengthened the critiques of AI voiced by sceptics and ultimately contributed to a significant decline in funding and institutional support. A paradigm shift began towards the end of the decade – and more prominently during the 1970s – with the rise of knowledge-based systems, a broad category encompassing various approaches that sought to represent and manipulate explicit knowledge.

The core idea was the ability to derive new knowledge from existing facts through deductive reasoning. In the subsequent decades, the capabilities of AI expanded considerably, with increasing attention to non-deductive mechanisms such as induction, analogy, and reasoning under uncertainty. These approaches gradually found their way into practical systems, contributing to the renewed growth of the field and its funding (Buchanan, 2005).

4.2.2 An overview of the key definitions of AI from the spring and summer period

Before delving into the individual milestones, it is useful briefly to summarise the context in which the first definitions of AI emerged. Each definition reflected not only the state of technology at the time, but also the prevailing scientific optimism, philosophical assumptions, and societal expectations. The following subsections highlight several representative formulations from the early "spring and summer" of AI, showing how researchers attempted to capture the essence of intelligence in machines through different perspectives and approaches.

Founding proposal and conference for initiation of AI studies; 1955

The early pioneers of artificial intelligence, working within the conceptual framework of their time, envisioned AI as "Strong AI" – the idea that machines could eventually replicate the mental processes, cognitive abilities, and functions found in the human brain. Definitions grounded in the notion of Strong AI were deeply intertwined with the field's foundational aspirations, even though the hardware resources available at the time were severely limited.

In part, this was due to a significant underestimation of the complexity of the human brain. As a result, the Strong AI concept remained largely philosophical – more a thought experiment or projection of scientific optimism than a realistically implementable vision.

This scientific optimism is exemplified by McCarthy et al. (1955), when they state that: "…every aspect of learning or any other features 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 … the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving" (ibid., p. 12).

AI based on deductive reasoning; 1959

In his work, John McCarthy (1959) – as a key figure in the early development of AI – examines computer logic operations and methods for representing information within machine memory. He argues that the realisation of AI crucially depends on the ability to reason with common sense, framing AI primarily as a system for deductive reasoning. He states that "A program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows" (ibid., p. 1).

Definition based on general intelligence; 1969

This definition by Minsky (1968) is still one of the most frequently cited ones and points in a simple and understandable form to the essence of the issue. Unfortunately, in the early 21st century, when human society is essentially surrounded by AI mechanisms, this definition is too narrow, especially for legislative, commercial, regulatory, or ethical purposes. It fails to account for the fact that many AI systems now perform tasks that no longer require human intelligence as a result of automation, and that their societal impact often lies not in how intelligent they are, but in how they are deployed, governed, and interpreted. AI is "the science of making machines do things that would require intelligence if done by men" (ibid., p. V).

Representation of knowledge; 1976

Allen Newell and Herbert Simon are undoubtedly among the most influential pioneers in the history of artificial intelligence. They entered the annals of the field with the development of the *Logic Theory Machine* (1956) and the *General Problem Solver* (1959). The latter is often considered the first significant approach to knowledge representation. Their definition of intelligent behaviour incorporates key concepts such as the real-world context, system goals, environmental adaptation, and complexity – making it particularly suitable for non-trivial applications, regardless of whether the domain in which the system operates is narrow or broad.

As they put it: "By 'general intelligent action' we wish to indicate the same scope of intelligence as we see in human action: that in any real situation, behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity." (Newell & Simon, 1975, p. 10).

4.3 AI Autumn: The Harvest

One of the reasons for the initial lack of progress in AI and the fact that the field

was almost defunded was that reality fell far short of expectations, often fuelled by grandiose "PR" proclamations and media "hype". This shows how the concept of AI can be misleading in that it leads to unrealistic social expectations and other negative connotations. Therefore, we may encounter the replacement of the term *AI* in the winter and early autumn of the field with alternative terms such as "computer science" or "analytics".

4.3.1 Expert systems and the end of the first wave

The reasons for this early direction in the development of AI can be traced to the nature of the first systems, such as ELIZA and the General Problem Solver (GPS), which aimed to imitate human intelligence and brain functions. These systems laid the groundwork for what became known as expert systems – AI applications characterised by a robust, rule-based approach that relied on domain-specific knowledge in narrowly-defined areas of expertise. The renewed optimism of the 1980s, often referred to as the second AI boom, centred around expert systems and the rational agent paradigm.

As noted by Toosi et al. (2021), expert systems quickly found practical applications in corporate settings, where they were employed for tasks such as market forecasting, spectrographic analysis, syllabus database search, and the evaluation of medical data. These expert systems were developed under the assumption that human intelligence could be formalised and reconstructed through algorithms following a top-down approach – the designers encoded all the necessary information into the system to fulfil predefined goals. Scientists attempted to simulate intelligence using rule-based systems grounded in formal logic, particularly binary operations such as "if-then" (implication) or "either-or" (disjunction).

However, this approach quickly encountered limitations, primarily because of the restricted computational capacity of contemporary hardware. More fundamentally, it was conceptually flawed: expert systems achieved promising results only in narrowly-constrained domains where knowledge could be explicitly codified in binary logic. In broader, more dynamic environments, this formalist paradigm proved inadequate – a point emphasised by Haenlein and Kaplan (2019), who highlight the limited adaptability of such systems in real-world contexts where uncertainty and nuance prevail.

The philosopher Hubert Dreyfus had already anticipated these shortcomings in his seminal critique *What Computers Can't Do* (1972), in which he argued that human intelligence is deeply contextual, intuitive, and embodied – characteristics that resist formalisation into discrete rules. The expert system model, he contended, relies on an overly simplistic and decontextualised view of cognition.

Consequently, a more flexible approach had to emerge – one that could interpret external data, learn from experience, and adapt dynamically to environmental changes. This more general form of AI marked a new phase of develop-

ment, in comparison to which the early expert systems appear primitive, rigid, and obsolete (Haenlein & Kaplan, 2019).

Initially, the field once again experienced a surge of exaggerated ambitions and unrealistic expectations regarding the imminent development of a fully general artificial intelligence. However, when these aspirations were tempered by the practical successes of expert systems – which, despite their limitations, demonstrated impressive performance within narrowly-defined domains – AI research gained new momentum.

This pragmatic turn allowed the field to grow again, particularly as AI applications began to show clear economic benefits and efficiency gains. These developments ushered in a new wave often described as the "summer" of artificial intelligence. A widely recognised milestone of this period is the 1997 victory of IBM's Deep Blue over the chess grandmaster Garry Kasparov.

While frequently cited as a breakthrough in machine intelligence, it is important to note that Deep Blue was not an example of deep learning. Rather, it was based on a combination of brute-force search, sophisticated evaluation functions, and pruning techniques – a form of symbolic AI or GOFAI. The system relied on massive computational power and expert-crafted heuristics, rather than learning from data, as modern deep learning systems do. Thus, while historically significant, Deep Blue represents a different lineage in the development of AI from the one that led to contemporary advances in artificial neural networks and deep learning architectures.

4.3.2 An overview of the key definitions of AI from the autumn period

During the "autumn" stage of the development of AI, the field moved beyond the initial enthusiasm and symbolic approaches, gradually consolidating into more formalised and pragmatic frameworks. Definitions from this period often reflect a turn towards rationality, agency, and measurable behaviour, emphasising the design of systems that can act purposefully within their environments. The following examples illustrate how the idea of AI was reframed in the 1990s in terms of rational agents and computational models of intelligent behaviour.

AI systems as rational agents; 1995

Russell and Norvig (2003) categorise definitions of AI into four foundational approaches: *systems that think like humans*, *think rationally*, *act like humans*, and *act rationally*. Each of these categories is illustrated with representative definitions drawn from key figures in the history of AI.

In the first category – systems that think like humans – they cite John Haugeland's 1985 definition: "The exciting new effort to make computers think… machines with minds in the full and literal sense." Similarly, Bellman (1978) characterises AI as "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning…".

In the second category – systems that think rationally – they include the definition by Charniak and McDermott (1985): "The study of mental faculties through the use of computational models" and that of Winston (1992): "The study of the computations that make it possible to perceive, reason, and act."

The third category – systems that act like humans – is illustrated by Ray Kurzweil's (1990) formulation: "The art of creating machines that perform functions that require intelligence when performed by people", along with the definition by Rich and Knight (1991): "The study of how to make computers do things at which, at the moment, people are better."

Finally, the fourth and most contemporary category – systems that act rationally – reflects a shift toward formal models of intelligent behaviour. Here, Russell and Norvig include the definitions by Poole et al. (1998): "Computational intelligence is the study of the design of intelligent agents" and Nilsson (1998): "AI is concerned with intelligent behavior in artefacts." This approach, emphasising rational agency through a blend of mathematics and engineering, is also reflected in Russell and Norvig's own working definition, which they present in their textbook: "We define AI as the study of agents that receive percepts from the environment and perform actions, and we cover different ways to represent these functions, such as production systems, reactive agents, real-time conditional planners, neural networks and decision-theoretic systems." (Russell & Norvig, 2003).

5. Discussion

Although this phase of the development of AI can be metaphorically described as the *autumn* of the life cycle – characterised by maturation, the consolidation of methodologies, and the refinement of rational agent architectures – it did not culminate in a full *AI winter*, as might have been expected by analogy. Rather than entering a period of stagnation or decline, the field underwent a *paradigmatic shift*, marked by the advent of systems that transcended the limits of symbolic or rule-based approaches. This shift inaugurated a *new spring*, driven by the emergence of data-driven models, particularly artificial neural networks and large-scale machine learning architectures, which began to spread rapidly across domains. These systems not only redefined the boundaries of what AI could achieve, but also became deeply integrated into existing sectors, transforming them from within and enabling entirely new capacities – as we can observe today in language processing, image recognition, autonomous systems, and decision-making support.

Today, the question is no longer whether AI will play a role in any sphere of society. As noted by Haenlein and Kaplan (2019, p. 6), the question is now posed in a completely different way: "The question is less whether AI will play a role in these elements but more which role it will play and more importantly how AI systems and humans can (peacefully) coexist next to each other. Which decisions

should rather be taken by AI, which ones by humans, and which ones in collaboration will be an issue all companies need to deal with in today's world and our articles in this special issue provide insights into this from three different angles." These are the questions that all institutions, companies, and individuals must solve in today's world.

5.1 Big Data and the Third AI Boom

After the first and second AI booms – the eras of symbolic and expert-system approaches – the current and third AI boom is defined by the rise of data-driven machine learning and deep neural architectures. By the late 1990s, the development and rapid expansion of powerful software and hardware capabilities – along with the growth of global data infrastructures and the internet – enabled the accumulation of immense datasets and the computational capacity required to process them. This shift directly accelerated further advances in the field of artificial intelligence.

The *third AI boom* – the new spring – was preceded by a notable change in how the field conceptualised itself: researchers became more conservative and sought to establish a more credible and methodologically grounded identity for AI – refining its subjects of study, research goals, and application domains. Among the credible and widely-adopted methods were statistical approaches such as *Hidden Markov Models* (HMM), which relied on rigorous mathematical principles and training on large real-world datasets. This methodological shift coincided with the emergence of well-structured public data repositories, compiling information from a wide range of sources. Together, these developments opened the door to a wide array of new application domains – including *text prediction, image analysis, handwriting recognition, voice processing, or bioinformatics*. These technologies rapidly found integration into industrial and commercial environments, supporting tools and systems across countless professions.

It is important to recognise that the current boom in artificial neural networks builds fundamentally on research dating back to the 1940s. A key figure here is the Canadian psychologist Donald Hebb, who proposed a theory of synaptic plasticity now known as *Hebbian learning* – the idea that the connection between two neurons is strengthened when they are activated simultaneously. This principle inspired the development of early neural network models that attempted to simulate certain features of biological neurons, such as weighted connections, threshold activation, and distributed parallel processing – albeit in highly simplified, abstracted form.

The development of artificial neural networks stalled in the late 1960s following critical findings by Marvin Minsky and Seymour Papert in their book *Perceptrons* (1969), where they highlighted the limitations of single-layer networks. At that time, the available computers lacked the processing power and memory required for more complex, multi-layered architectures.

With the resurgence of computational power and the availability of massive datasets, artificial neural networks have returned with renewed strength under the label *Deep Learning*. As Haenlein & Kaplan (2019) note, this approach has been central to recent breakthroughs – such as *Google's AlphaGo*, which succeeded in defeating top-ranked human players in the game of *Go*. Go is significantly more complex than chess; while chess begins with 20 possible moves, Go starts with 361. This accomplishment marks another historic milestone, following Deep Blue's victory over Garry Kasparov, and illustrates the leap in computational strategy enabled by deep neural architectures.

At the turn of the millennium, hardware capabilities underwent another transformative leap. As early as 2006, graphical processing units (GPUs) were shown to be approximately four times faster than traditional central processing units (CPUs) in executing neural network operations – a gap that widened dramatically, with GPUs becoming up to 60 times faster by 2011; Ciresan et al. (2011) demonstrated a very significant acceleration of GPU implementations of deep CNNs compared to CPUs. Interestingly, this technological acceleration was largely propelled not only by academic or industrial AI demand, but primarily by the rapidly-growing *digital gaming industry*, which had previously been dismissed as mere entertainment for children and adolescents.

According to Toosi et al. (2021), this surge in processing power enabled machine learning algorithms to outperform human experts in increasingly specialised domains. Over time, these capabilities expanded across virtually all sectors of society, transforming AI from a niche tool into an integral and ubiquitous component of modern technological systems.

Since AI systems based on data-driven, statistically optimised learning models – significantly boosted by the availability of affordable and massively parallel GPU hardware – have become an integral part of the reality we experience, definitions of AI must evolve accordingly. First, they need to address the phenomenon of massive datasets, which may include extremely sensitive personal information – often collected without informed consent, or with consent granted unknowingly. This raises serious concerns about privacy violations, as well as the potential for highly effective manipulative techniques or even criminal activities such as blackmail.

Furthermore, modern definitions must also account for the multitude of application layers and domains in which AI is deployed. What is valid and appropriate for medical AI applications, for example, may be entirely unsuitable for legislative, scientific, educational, or entertainment contexts.

This leads us to a conceptual impasse: either we adopt many mutually incompatible definitions, each tailored to the needs and norms of its respective domain, or we attempt to enforce a single universal definition – one that may be too narrow to be inclusive, too broad to be meaningful, or so vague that it obstructs innovation and regulatory clarity. In other words, the

paradox remains: any definition of AI is likely to be too narrow, too broad, or both at once.

5.2 AI as a Machine Learning Phenomenon and an Institutional Perspective

The principle of broad-narrow definitions of AI is not just a metaphor, but a completely practical problem. Different political, business, national and transnational institutions and alliances are characterised by different agendas, focus, operations, and management and therefore different definitions and definitional frameworks are appropriate for them. And every term in the sentence of the definition is fundamentally important.

An example can be the ongoing debate on the definition of AI within the OECD. The current definition of AI within the OECD is the following: "An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy." (Perset & OECD, 2023, p. 4).

The newly proposed definition in 2023 reads as follows: "An AI system is a machine-based system that can, for explicit or default objectives, generate outputs such as predictions, recommendations, or decisions that influence physical or virtual environments. AI systems are designed to operate with varying levels of autonomy." (Perset & OECD, 2023, p. 4).

At first glance, it may seem that initiating the legislative and expert commentary processes across many countries and institutions over "a few words" is unnecessary. However, the opposite is true – because these *few words* significantly shift the meaning and potential legal implications of the definition.

Take, for instance, the change from *"influencing real or virtual environments"* to *"that influence physical or virtual environments"*. While the change appears minimal, it reflects a more deterministic view of AI systems' impact. The original phrase "influencing" implies a possibility or potential to affect the environment, whereas "that influence" asserts a more direct, active role of AI systems in shaping outcomes. This change could affect legal interpretations of responsibility, liability, and the scope of regulation. Similarly, replacing "real" with "physical" introduces a more precise, perhaps legally or technically grounded, term. The shift from "real" to "physical" may seem subtle, but it reflects a significant conceptual narrowing. While "real" encompasses both physical and virtual environments with tangible consequences, "physical" refers strictly to the material, natural world. In an era in which digital experiences can have real emotional, social, and economic impacts, this change implicitly excludes a growing domain of reality that is no less consequential.

It may seem like a simple idea – even a minor linguistic distinction – that we are explaining at length. However, this distinction is crucial, as even a single word

can significantly reshape the scope of a legal or conceptual definition. The term *real* has evolved in meaning during the digital age. Today, *virtual reality* is not just a metaphor but a lived experience, integrated into everyday life through digital platforms, smart devices, and wearable technologies.

These developments blur the boundaries between physical and digital environments. An action taken in a virtual context – such as a financial decision, social interaction, or personal disclosure – can have tangible and lasting effects in a person's physical world. As Toosi et al. (2021) argue, the consequences of digital actions are increasingly real in terms of their psychological, social, and ethical impact.

Therefore, it is necessary to move beyond a simplistic dichotomy of *real* versus *virtual*. Rather than dismissing one as artificial and the other as essential, we must acknowledge that both domains generate effects that are real in terms of lived human experience. This calls for a more nuanced philosophical reflection on what constitutes *reality* today – especially in the context of AI systems that shape our choices, perceptions, and sense of self. Importantly, this use of the term *multi-layered reality* refers not to computational architectures, but to the layered structure of human experience across physical, social, and digital dimensions.

Definition in the legal and political--economic framework of the European Union

The complexity of developing a definitional framework that will impact on hundreds of millions of lives is illustrated by the process of formulating a definition of AI acceptable to the entire European Union. This definition can be found in Regulation (EU) 2024/1689, known as the *AI Act*, which was adopted by the European Parliament on 13 March 2024 and by the Council on 21 May 2024. It entered into force on 1 August 2024, and its provisions were due to be phased in over the months that followed. The definition is as follows: "'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (AI Act, p. 46).

The debate regarding definitions within the political-economic domain of the EU is still alive, in constant change, and difficult to grasp conceptually. Nevertheless, this section outlines the development of key documents and activities during the selected periodand the development process of the definition.

The European Commission document from 2018 used the following definition of AI: "Artificial intelligence is considered to be systems exhibiting intelligent behaviour in the form of evaluating their surroundings and subsequently making decisions or taking steps – with a certain degree of autonomy – to achieve specific goals. Systems using artificial intelligence technology can be purely software-based,

operating only in the virtual world (e.g., voice assistants, image analysis programs, search engines, voice and face recognition systems), or they can be built into hardware (e.g., advanced robots, autonomous vehicles, drones and various forms of use of the Internet of Things)." (European Commission, 2018, p. 1)

This definition was further developed a year later by the High-Level Expert Group on Artificial Intelligence (HLEG AI), which proposed the following revised version: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information derived from this data and deciding the best actions to take to achieve the given goal. AI systems can also be designed to learn how to adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, artificial intelligence encompasses several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are concrete examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)." (HLEG AI, 2019b, p. 6)

However, this definition has certain limitations and should not be treated as all-encompassing. For instance, the phrase "systems designed by humans" may not adequately capture recursive or emergent forms of AI design, where systems are generated by other AI systems – even if these are, ultimately, traceable to human-originated architectures.

This definition is further elaborated, analysed, and expanded in a broad context in 2020 through the EU JRC (Joint Research Centre) scientific hub report *AI Watch: Defining Artificial Intelligence* (Samoili et al., 2020). The above-mentioned definition (HLEG AI, 2019b) is called the baseline definition by the researchers in this analysis. The JRC document provides a detailed taxonomic analysis of many definitions, whether legislative and political-economic or definitions produced by world-renowned scientific authorities over time as AI has developed, which are called operational definitions. The operational definition is also co-created by a set of keywords that characterise the basic and cross-sectional areas of AI – each of the definitions is analysed and assigned to a specific domain (e.g. reasoning, planning, learning, or ethics and philosophy...) and keywords that define it.

Dozens of definitions are sorted here according to domains (e.g. EU, non-EU, international organisations, research purposes...), commented on, and put into context and are assigned to the fields from which they originated (e.g. academia, medicine, industry...). In order to achieve a common understanding of the concept

of AI within *AI Watch*, it is important that the starting point is the already-mentioned inclusive definition of HLEG – the baseline definition (HLEG AI, 2019b). According to the authors, it includes all technological developments and activities carried out by all types of actors that make up the AI ecosystem, whether these are industrial, research or government initiatives. The entire study thus serves as a map to navigate the maze of a large number of definitions relevant to selected sectors (Samoili et al., 2020).

In 2021, the European Union published the first draft law on artificial intelligence (the AI Act). In 2021–2022, the initiatives of the European Commission according to one of the sub-documents (SWD/2021/84, 2021) were supposed to lead to:

- the adoption of a European legal framework for AI
- the clarification of European rules on the new challenges regarding responsibility arising from these new (disruptive) technologies
- revision of existing sectoral security legislation.

The aforementioned European AI legal framework is built on initiatives providing:

- a definition of AI systems
- a definition of high-risk AI systems
- shared rules according to which AI services will be considered trustworthy in the EU market.

However, the document explicitly draws attention to the issue that the current EU law does not establish a definition of an AI system or horizontal rules for its use regarding the classification of risks associated with AI technologies. An important question is whether it is harmful or beneficial to look for and introduce a basic definition or an operational type of definition.

In another document related to the AI Act (European Commission, 2021) a single, unifying baseline definition of AI is proposed for the purposes of European law. The intention is for it to be future-proof, taking into account the rapid and unpredictable technological development and changes in the market for AI technologies, and at the same time to be sufficiently technologically neutral. Key participants in the AI value chain are also clearly defined here. This definition reads: "… 'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (ibid., Art. 3).

The AI techniques and approaches listed in the aforementioned Annex I are: "a) machine learning approaches, including tutored, untutored and reinforcement learning, using a variety of methods, including deep learning; b) logic and knowledge-based approaches, including knowledge representation, inductive (logical) programming, knowledge bases, inferential and deductive mechanisms, (symbolic) reasoning and expert systems;

c) statistical approaches, Bayesian estimation, search and optimization methods" (European Commission, 2021).

Definition efforts under the AI Act have been long-standing and are still ongoing, leading to the adoption of "provisional definitions" such as the GPAIS (General-Purpose Artificial Intelligence Systems) definition adopted in May 2023 by the EU Parliament. The Slovenian EU Presidency defined this technology as an "artificial intelligence system ... capable of performing generally applicable functions such as image/speech recognition, audio/video generation, pattern detection, question answering, translation, etc." (Council of the European Union, 2021). The French EU Presidency further emphasises that GPAIS "can be used in many contexts and can be integrated into many other artificial intelligence systems" (Council of the European Union, 2022a). Updated versions of the proposal discuss the role of actors in the development and value chain of the GPAIS (Council of the European Union, 2022b; Gutierrez et al., 2023). As can be seen in the cited documents, non-political institutions and experts under the heads of organisations such as the FLI (Future of Life Institute) are also involved in efforts to achieve a consensual and functional definition working on a clearly defined difference between the definition of a narrow (fixed) AI and a general AI definition (GPAIS). This division can be more precise and less exploitable and manipulable than the original division into four risk categories (unacceptable, high, limited, and low risk).

In the comment procedure of 2021, the panel of experts of the Association of Industry and Transport of the Czech Republic as a member of BusinessEurope states that "The definition of artificial intelligence, which is to be regulated in the document, is problematic – it is too broad and too narrow" (SPCR, 2021). Note: the document is not publicly available.

The topic has already been explored, but let us dwell on this claim one more time. Although this opinion sounds banal, it is a fundamental finding and reminder. It must be added that it is actually both too broad and too narrow and at the same time it is adequate; this is a practical demonstration of breaking Aristotle's analytic law of the exclusion of the third! As already mentioned, the definition depends on the purpose of its use and the goals of the person using it. And as the purposes and goals change with the needs and interests of different stakeholders, so does whether the definition is narrow, broad, or adequate.

This relativising element needs to be reckoned with and accommodated, as it cannot in principle be eliminated! Therefore, any unifying and exclusive baseline definition will inevitably suffer from a certain degree of discord – what might be called a definitional *trinity* of being simultaneously too broad, too narrow, and yet still potentially appropriate. As previously noted, the adequacy of a definition depends on the specific goals and intentions of those who use it. These goals vary across different stakeholders and evolve over time.

Consequently, no single definition can fully satisfy all contexts. Only a set of multi-layered, officially recognised, and legally grounded operational definitions – each tailored to particular needs – can provide the flexibility required to ensure that, within a given operational framework, a definition may be considered sufficiently adequate.

In other words, we must accept that definitions of AI are inherently operational and contingent: not universally adequate, but potentially appropriate within predefined contexts, interests, and regulatory needs.

6. Conclusion

While the historical and conceptual overview has shown how definitions of AI evolve in relation to scientific progress and societal needs, the conclusion must also reflect the broader consequences of these definitional choices. Beyond technical accuracy, definitions influence ethical debates, legal frameworks, and social perceptions. Therefore, the closing part of this study turns to the ethical and social implications, highlighting how the very act of defining AI carries a fractal dimension that affects multiple layers of society.

6.1 Ethical and Social Issues – the Fractal Dimension of the Definition

The discussion on the ethical and social issues associated with different definitions of AI and their impact on society can be conducted very broadly, as numerous interrelated elements and variables play a crucial role. One way to illustrate this complexity is by examining the inherent unknowability of how complex neural networks and deep learning algorithms function.

These self-learning systems are both large-scale and non-transparent. Because of their sheer scale and architectural complexity, it is often time-consuming – and in some cases practically impossible – to trace how they arrive at their outputs, even when their performance appears effective in practice. This is known as the black box phenomenon, where both the input and output data may be transparent and controlled, but the internal logic of the algorithm remains opaque and inaccessible. We previously encountered deep learning in the context of the game of Go, where this lack of transparency is ethically unproblematic. However, the situation changes dramatically when the same algorithmic structure is applied to sensitive content – for instance, when the input is medical data and the output is a diagnosis, prognosis, recommendation with life-altering consequences, or legal advice.

The ethical risk increases further when we consider biases – both those replicated from training datasets and those stemming from human cognitive tendencies to overtrust algorithmic decisions. This creates a deeply problematic dynamic. As Zemčík (2020) argues, the issue becomes morally explosive when biases are no longer limited to human

actors but are embedded in the "digital intellect" of autonomous virtual agents. Although several authoritative sources – such as the Ethics Guidelines for Trustworthy AI (HLEG AI, 2019a) and DARPA's Explainable AI (XAI) roadmap (Gunning & Aha, 2019) – address these issues in detail, there is still no unified consensus on how to mitigate such algorithmic externalities or eliminate the moral hazards they pose.

The situation observed in previous phases of the development of AI (often described as the spring, summer, autumn, and winter of AI), which has already been noted, is once again repeating itself – particularly in the form of unrealistic media representations and exaggerated claims by some experts. According to a 2019 report by Vincent, approximately 40% of startups in Europe that claim to use AI in their services actually do not (Vincent, 2019). This highlights the ongoing ambiguity surrounding the definition of AI – not only for the media and the general public, but even for professionals in the field.

Another ethically problematic practice is the exploitation of symbolic power and media hype associated with AI. Given the steady growth of investment in this sector, it is evident that some companies – either deliberately or unknowingly – are attempting to position themselves within this trend by misusing terms such as *AI* or *deep learning* or applying them loosely. This ambiguity underscores the need for clearer definitions, as well as for a well-structured ecosystem that reflects the interconnection of related disciplines and accounts for their specific needs within a multi-level definitional framework (Toosi et al., 2021).

6.2 A Definition to Eliminate the Cognitive Bias of Anthropomorphising AI Algorithms

If we consider the illusory side of artificial intelligence, which causes in people an ecstatic need to pay attention to virtual entities and communicate with them "as if they were conscious, living intelligences", we find that this illusion may be at the core of many – from the ethical (but also legal, economic, security...) point of view – questionable cases of the misuse of algorithms. Let us just give an outline of this issue to justify a new perspective on the definition of AI, which aims to make users aware of the automatism in their thinking of "attributing higher skills to AI" than it actually possesses. Users frequently attribute disproportionately more competences and typically human skills to AI than it wields – thus anthropomorphising and humanising AI.

It is astonishing how one of the most primitive chatbots, ELIZA, written in 1964–1966, which had the task of playing a word game with the user in such a way as to act like a psychotherapist, confused people in this way. Users entrusted it with their sensitive secrets after the exchange of a few lines, regardless of what would happen to this data, where it would be stored, how it would be handled, and whether it could ultimately be used against them.

Today, chatbots such as China's Dubao, Xiaoice, DeepSeek, or Tencent's Yuanbao are already important social media influencers, songwriters, writers, and many users consider them intimate friends or even choose them as life partners! Chatbots are capable of causing a global media scandal when users mistake the text output they generate, guided by the statistical evaluation of past conversations, for the prophecy of a hyper-intelligent entity or the morally corrupt behaviour of a conscious being – as was the case with the TAY chatbot. ChatGPT is frequently used inappropriately instead of professional databases, and users only later find out that the information provided was not facts, but strings of words intended to make a credible impression on the given user in the given domain. While newer versions have improved at providing sources and references, linking to a source is not a universal solution – it still raises critical issues such as the relevance of the citation, its context, and whether it truly supports the claim that is generated.

And this is not limited to chatbot-type AI. Every AI system, even when it interacts solely with other virtual entities, contributes to the creation of an environment that can be subjectively experienced by humans. This experience requires interpretation, the assignment of meaning, and integration into a network of familiar relationships – which itself constitutes a form of communicative interaction.

When users interact with a chatbot and get the impression that the AI displays certain traits of "human intelligence", they are, in fact, responding to carefully engineered illusions – what could be described as an *imitation game* or a *textual credibility simulation*. These effects are intentionally crafted by developers to evoke human-like responses. As Zemčík (2020) argues, this often leads to an unconscious tendency in users to treat the AI as if it were endowed with human consciousness and moral agency. We can illustrate the profound impact of anthropomorphising AI it in the widely-known case of a 14-year-old boy, Sewell Setzer III from Florida. He developed an emotional attachment to an AI character modelled as Daenerys on the Character.AI platform and later died by suicide. This tragedy underscores how users may perceive AI systems as conscious and emotionally responsive agents, leading to serious psychological and ethical implications (Blake, 2024).

It can feel as though the algorithm "experiences" the world – and not only that, but that it does so in the same way as humans. It is as if AI were part of the symbolic order (legal, moral, political, or socio-economic), which is traditionally reserved for conscious beings. But this software is not accountable for what it says. It has no past or future to reflect on, no capacity to "corrupt its soul" or "damn its perspective" – and likewise, no ability to derive satisfaction from doing good.

Therefore, it is important to be aware of these cognitive biases in certain situations. Under these circumstances, the partial goal of the definition (or the final one) should be preventive, protective, and pedagogical – to protect and educate

the user, show them their limitations, and provide them with the competences and skills to eliminate their own undesired thought automatisms, biases, to the greatest extent possible! And for that reason, the very use of the term "artificial imitation of intelligence", "imitation game", or "text credibility game" alone, which in these cases can replace the usual term "artificial intelligence", can be very helpful. In the body of the definition itself, it is then necessary to emphasise that it is an illusory intelligence created for the purpose of acting in a certain way. It needs to be emphasised that it is quite transparently the intention of the creators of the AI that users have a tendency to attribute properties to the algorithm (or behave in that way in its environment) that it does not have. In this way, AI can be used against the interests of users as well as the interest of society – all as a result of detailed knowledge of the workings of the human psyche and attention, accelerated by an immense amount of training data and computing power.

Use of Artificial Intelligence Tools: The author declares that generative artificial intelligence tools (ChatGPT, OpenAI; various versions) were used during the preparation of this manuscript exclusively for supportive purposes, including consultation on the structure of the text, brainstorming of ideas, assistance with literature orientation, support in drafting and organising summary tables, and consultation on the translation and clarification of selected technical terms into a foreign language.

The AI tools were not used to generate original scientific content, data, analyses, or interpretations, nor to replace critical scholarly judgment.

The author bears full responsibility for the content of the article, including the accuracy of citations, interpretations, and conclusions.

References

Aristotle. (n.d.). *Topics.* http://classics.mit.edu/Aristotle/topics.html

Aristotle. (1905). *Aristotle's Politics.* Oxford: Clarendon Press.

Barták, R. (2017). *Co je nového v umělé inteligenci [What's new in artificial intelligence].* Nová beseda.

Bellman, R. E. (1978). *An introduction to artificial intelligence: Can computers think?* Boyd & Fraser.

Berglund, L., Stickland, A. C., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., … Evans, O. (2023). Taken out of context: On measuring situational awareness in LLMs. *arXiv*, pp. 1–41.

Blake, M. (2024, October 23). Mother says AI chatbot led her son to kill himself in lawsuit against its maker. *The Guardian*. https://www.theguardian.com/technology/2024/oct/23/character-ai-chatbot-sewell-setzer-death

Bostrom, N. (2016). *Superintelligence*. Oxford University Press.

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *AI Magazine, 26*(4), 53–60.

Charniak, E., & McDermott, D. (1985). *Introduction to Artificial Intelligence*. Addison-Wesley.

Ciresan, D., Meier, U., Masci, J., Gambardella, L. M., Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2, 1237–1242.

Council of the European Union. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts – presidency compromise text. https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf

Council of the European Union. (2022a). Proposition de Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union – Text de compromis de la présidence – Article 3, paragraphe 1 ter, Articles 4 bis à 4 quater, Annexe VI (3) et (4), considérant 12 bis bis. https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-FRA-Art-34-13-May.pdf

Council of the European Union. (2022b). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts – General approach. https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. Harper & Row.

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

European Commission. (2018). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe (COM 2018/237 final). European Commission, 25 April 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0237

European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM 2021/206 final). European Commission, 21 April 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Feigenbaum, E. A., & Feldman, J. (1963). *Computers and thought.* McGraw-Hill.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality.* Oxford University Press UK.

Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI Magazine, 40*(2), 44–58.

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C., & Franklin, M. (2023). A proposal for a definition of general purpose artificial intelligence systems. *Digital Society, 36*(2).

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review, 61*(4).

HLEG AI. (2019a, April 8). *Ethics guidelines for trustworthy AI.* From an official website of the European Union: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

HLEG AI. (2019b). *A definition of artificial intelligence: Main capabilities and scientific disciplines.* European Commission, High-Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–1554. https://www.cs.toronto.edu/~hinton/absps/ncfast.pdf

Johnson, K. L., Kruger, J., Schlösser, T., & Dunning, D. (2013). How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning–Kruger effect in self-evaluation of performance. *Journal of Economic Psychology, 39*, 85–100.

Kurzweil, R. (1990). *The age of intelligent machines*. MIT Press.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1–46. http://vision.stanford.edu/cs598_spring07/papers/Lecun98.pdf

McCarthy, J. (1959). *Programs with common sense.* Stanford University. http://www-formal.stanford.edu/jmc/

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence.* Stanford University. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Mikolov, T., Joulin, A., & Baroni, M. (2018). A roadmap towards machine intelligence. *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016,* Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part I 17, pp. 29–61. Springer.

Minsky, M. (1968). *Semantic information processing.* MIT Press.

Minsky, M., & Papert, S. (1969). *Perceptrons.* MIT Press.

Newell, A., & Simon, A. (1956). The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory, 2*(3), 61–79.

Newell, A., & Simon, A. (1959). General Problem Solving program for a computer. *International Conference on Information Processing,* Paris, June 13–20, pp. 10–16. https://iiif.library.cmu.edu/file/Newell_box00039_fld03042_doc0001/Newell_box00039_fld03042_doc0001.pdf

Newell, A., & Simon, A. (1975). Computer science as empirical enquiry: Symbols and search. *ACM Turing Award Lecture. Carnegie-Mellon University; December 1, 1975.* https://iiif.library.cmu.edu/file/Newell_box00024_fld01660_doc0003/Newell_box00024_fld01660_doc0003.pdf

Nilsson, N. J. (1998). *Artificial intelligence: A new synthesis*. Morgan Kaufmann.

Perset, K., & OECD. (2023). *Preliminary thoughts on possible updates to the definition of an AI system contained in the OECD AI Principles*. OECD AI Policy Observatory. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf

Poole, D. L., Mackworth, A. K., & Goebel, R. (1998). *Computational intelligence: A logical approach.* Oxford University Press.

AI Act. (2024). Regulation (EU) 2024/1689. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Rich, E., & Knight, K. (1991). Artificial Intelligence (2nd ed.). McGraw-Hill.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach.* (2nd ed.). Prentice Hall.

Samoili, S., Montserrat, L. C., Emilia, G. G., Giuditta, D. P., Fernando, M.-P., & Blagoj, D. (2020). *AI WATCH: Defining artificial intelligence.* Publications Office of the European Union.

Schuchmann, S. (2019). *Analyzing the prospect of an approaching AI winter.* (Bachelors Thesis). University of Liverpool.

SPCR. (2021). *Position of the platform for AI of the Confederation of Industry of the Czech Republic.* Svaz průmyslu a dopravy. https://www.spcr.cz/aktivity/evropske-a-mezinarodni-vztahy/sp-v-mezinarodnich-organizacich/businesseurope

SWD/2021/84, f. (2021, April 21). Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts: Commission Staff Working Document Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council. European Comission. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=SWD:2021:84:FIN

Tegmark, M. (2018). *Life 3.0.* Penguin UK.

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

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Vincent, J. (2019, March 5). Forty percent of 'AI startups' in Europe don't actually use AI, claims report. *The Verge*. https://www.theverge.com/2019/3/5/18251326/ai-startups-europe-fake-40-percent-mmc-report

Winston, P. H. (1992). *Artificial intelligence.* (3rd ed.). Addison-Wesley.

Zemčík, T. (2020). Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI & Society, 36*, 361–367.

Tomáš Zemčík
VSB – Technical University of Ostrava, Department of Social Sciences;
email: tomas.zemcik@vsb.cz

ZEMČÍK, T. Stručná historie definic umělé inteligence: od velké Bomby po Black boxy

Studie zkoumá a zdůrazňuje přímý vztah mezi dobovým věděním, paradigmaty, cíli a očekáváními veřejnosti v oblasti umělé inteligence (AI) a jejími definicemi. Pro účely tohoto výzkumu bylo pro rozdělení fází vývoje AI použito analogie k ročním obdobím: jaro až zima AI. Historie AI je zde pokryta pouze v rozsahu nezbytném k poukázání na tento vztah s možností odvození vyplývajících definic. Pro každé období jsou uvedeny příklady definic AI typických pro dané období. Tato historická exkurze je poté použita jako podklad pro úvahy o formě (nikoli obsahu) definice AI, která je vhodná pro současný stav oboru a jeho paradigma, zaměření a použití. Je nastíněna současná diskuse o podobě definice AI v rámci legislativy EU. Forma vhodné definice AI pro současnost je zkoumána z pohledu zainteresovaných aktérů, jako jsou nadnárodní subjekty, obchodní organizace a další zúčastněné strany, a je porovnána s některými již platnými definicemi těchto subjektů. Je poukázáno na paradox v definicích AI, které jsou z určitého hlediska vždy příliš „úzké a zároveň široké". Nakonec je v rámci koncepční ideace prozkoumána možnost použití fraktální definice s pevným racionálně-morálním jádrem, ale s měnícím se obsahem v závislosti na úrovni, na které je aplikována. Tato operativní fraktální definice by v zásadě mohla vyřešit stále přítomný paradox „širokosti–úzkosti".

Klíčová slova: *umělá inteligence, definice, algoritmy černé skříňky, historie, regulace a etika, kognitivní zkreslení*

Appendix

Table 1. Chronological overview of the development of AI through representative systems, concepts, and milestones

Period/ "Season"	Approx. Years	Representative Systems/ Events/Authors	Main Characteristics/ Definition Focus
Predecessors of AI	Antiquity–1940s	Homer's *tripods*; Aristotle's "soulless tools"; Descartes' automata; Leibniz's reasoning machines; Čapek's *R.U.R.*; Baum's *Tiktok*; Mechanical Turk	Cultural and philosophical imagination of artificial beings; metaphors for mechanical intelligence; "prosthetics of the mind" (Floridi, 2014)
AI Spring – The Birth of AI	1943–1956	**Alan Turing** (*Computing Machinery and Intelligence*, 1950); **Turing Test**; **Dartmouth Conference** (McCarthy et al., 1955)	Foundational stage; AI as "machines that think"; first operational test of intelligence; optimistic "Strong AI" vision
AI Summer – The Golden Age	1956–1969	**Newell & Simon** (*Logic Theory Machine*, *General Problem Solver*, 1956); **Minsky** (*Semantic Information Processing*, 1968); **Feigenbaum & Feldman** (*Computers and Thought*, 1963)	Symbolic reasoning and heuristic search; intelligence as formal manipulation of symbols; GOFAI paradigm; heuristic methods
AI Fall/ Autumn – From Expert Systems to Rational Agents	1970s–1990s	**ALPAC (1966)** and **Lighthill (1973)** reports; **Dreyfus** (*What Computers Still Can't Do*, 1992); **Expert systems** (e.g. MYCIN, DENDRAL); **Russell & Norvig** (*Artificial Intelligence: A Modern Approach*, 2003); **Kurzweil (1990)**; **Poole et al. (1998)**; **Nilsson (1998)**	Disillusionment with symbolic AI (first AI winter) followed by pragmatic recovery; emergence of knowledge-based systems; transition to rational-agent paradigm; AI defined as systems that "think or act rationally"
Big Data & The Third AI Boom – The New Spring	2000s–2010s	**LeCun et al**. (1998), **Hinton et al.** (2006); **Ciresan et al. (2011)**; **AlphaGo (2016)**; **GPT architectures (2018→)**	Data-driven and neural models; massive datasets and GPU revolution; intelligence as adaptive pattern recognition; beginning discussion of explainability and ethics
Contemporary AI – Regulation, Ethics and Fractal Definition	2020s–present	**OECD (2023)**; **EU AI Act (2024)**; **AI Watch (Samoili et al., 2020)**	Operational and legal definitions; focus on autonomy, adaptiveness, responsibility, and ethical transparency; recognition of "too narrow and too broad" paradox in AI definitions

Table 2. Overview of AI "booms" by period and characteristics

Boom	Approx. Years	Typical Label in Your Text	Main Focus/ Technological Paradigm	Key Authors/ Systems Mentioned	Outcome
First AI Boom	**mid-1950s – mid--1970s**	*AI Spring – Birth of AI → AI Summer – Golden Age*	Symbolic logic, rule-based reasoning, early heuristics (*GOFAI*)	Turing (1950); McCarthy et al. (1955); Minsky (1968); Newell & Simon (1956, 1959)	Great optimism and funding; collapse after ALPAC (1966) and Lighthill (1973) reports **→ First AI Winter**
Second AI Boom	**mid-1980s – late 1990s**	*AI Fall/Autumn – From Expert Systems to Rational Agents*	Expert systems, knowledge-based reasoning, later rational-agent models	Dreyfus (1992); Feigenbaum & Feldman (1963); Russell & Norvig (2003); Kurzweil (1990)	Renewed optimism via expert systems; ended with market saturation & high maintenance costs **→ Second AI Winter**
Third AI Boom	**2000s – present**	*Big Data & The New Spring*	Data-driven ML, Deep Learning, neural networks, LLMs	LeCun et al. (1998), Hinton et al. (2006); Ciresan et al. (2011); AlphaGo (2016); GPT series	Ongoing era; exponential growth; raises ethical, regulatory and societal questions (AI Act, 2024)