AI Paradigms and Agent-based Technologies

Luis Macedo

Abstract Artificial Intelligence (AI), the science and engineering of artificial intelligent agents, is a multidisciplinary and ubiquitous field that has become increasingly important to society. In fact, some argue that AI is at the heart of a new industrial revolution. However, with the rapid pace of research and development, the impact of AI on our lives is still unfolding. While there are certainly a myriad of positive impacts, there are also concerns that have led many experts to call for greater regulation and control. In this chapter, we address the main paradigms of AI and specifically focus on the perspective of artificial intelligent agents, a common denominator for all the paradigms and one of the most consensually accepted approaches for AI, without forgetting their interaction with human agents. We start by reviewing the history of AI and contextualizing its major paradigms over time, highlighting the key breakthroughs that have brought AI to its present state of grace. From there, we delve into the perspective of artificial intelligent agents and the various technologies that rely on them. We superficially examine issues related to coordination, cooperation, and negotiation between artificial and human agents, with special focus on human-AI agents cooperation. We emphasize the importance of viewing these agents as part of a broader ecosystem in which both humans and artificial agents "live" in a symbiotic relation. In the light of this ecosystem, where human-AI cooperation flourishes, we examine the human and machine in the loop concepts, more precisely, we discuss the various ways in which humans are involved in AI, including as sources of data for AI agents, beneficiaries of AI agents output, and in the design of AI agents (e.g., creation of knowledge representation structures, and coding decision-making and machine learning algorithms). We overview the fundamental areas of machine learning, including reinforcement learning and the related fields of learning by demonstration, apprenticeship learning, imitation learning, and inverse reinforcement learning, which hold great promise for the future of AI. Finally, we

Luis Macedo
Center of Informatics and Systems of the University of Coimbra, University of Coimbra, e-mail: macedo@dei.uc.pt

discuss the limitations of AI, the challenges AI faces, the opportunities that lie ahead, with all their positive and negative impacts.

1 Introduction

Artificial intelligence (AI) [56] is, on one hand, the science that seeks to study, comprehend and model forms of intelligence and various behavioral manifestations of intelligence (not necessarily human intelligence), and, on the other hand, it is the engineering practice of incorporating these forms of intelligence into machines. In doing so, AI serves as an instrumental tool to mimic, model, and augment not necessarily but mainly human intelligence. It is thus the science and engineering of artificial intelligent agents, i.e., agents that solve problems, perform tasks and make decisions that require intelligence. This definition of AI presupposes a dialectic between its facets of science and engineering, in which the results of each of the strands feeds the other. It is important to note that intelligence is a multidimensional concept, and different theories and perspectives may emphasize different aspects or components of intelligence. Most definitions emphasize cognitive abilities such as reasoning, problem-solving, learning, decision-making, and adaptability as core components of intelligence [66, 47, 27, 28, 37]. Intelligence is also often characterized by the ability to acquire, understand, and apply knowledge effectively to achieve goals or solve problems. It can manifest in various forms, such as linguistic, logical-mathematical, spatial, interpersonal, and others, according to different frameworks and theories. In any case, there appears to be a more expansive definition that encompasses the ability to achieve goals across diverse environments. [37].

Nowadays, AI is an ubiquitous and important multi-disciplinary field, providing help for many of the most challenging problems in many areas of the society. In fact, AI is progressively asserting its significance in our contemporary society [41]. Indeed, many assert that AI is the driving force behind a new Industrial Revolution, redefining the way we live and work. The rapid pace of research and development in AI has brought forth transformative changes. Yet, the full extent of AI's impact on our lives remains an ongoing narrative that began to be written long ago with the contributions of thinkers such as Aristotle, Babbage, and all the researchers involved in the early problem solvers of the 20th century. Amidst its myriad of affirmative consequences, concerns have emerged, prompting experts to advocate for more stringent regulatory measures and control mechanisms.

In this chapter, we undertake an in-depth exploration of AI paradigms, with a specific emphasis on artificial intelligent agents, a unifying perspective that spans across AI paradigms and stands as one of the most widely accepted facets of AI. Our analysis takes into account the intricate interplay between these artificial agents and their human counterparts. We commence by tracing the historical trajectory of AI, delineating the evolution of its primary paradigms over time, and spotlighting the pivotal breakthroughs that have propelled AI to its current state of eminence.

Venturing further, we delve into the world of artificial intelligent agents and the array of technologies reliant upon them. Our narrative underscores the imperative of considering these agents as integral components within a broader ecosystem in which humans play a pivotal role. We accentuate the adoption of best practices and methodologies that epitomize effectiveness, efficiency, and transparency. A central theme explored is the "human in the loop" concept, which underscores the fundamental role of humans in the dynamic interplay between artificial intelligence and human systems. We scrutinize the diverse ways in which humans are intricately woven into the fabric of AI, serving as sources of data, recipients of AI-generated outcomes, and architects of knowledge representation structures and decision-making algorithms.

This discourse also casts a spotlight on issues pertinent to coordination, cooperation, and negotiation between artificial and human agents. Notably, we delve into the bedrock of machine learning (ML), encompassing reinforcement learning and learning by demonstration, both of which hold immense promise for the future of AI.

In culmination, we cast our gaze on the limitations encumbering these methods, the formidable challenges that lie ahead, the myriad of opportunities ripe for exploration, and the conceivable positive and negative ramifications of AI. Our comprehensive analysis seeks to shed light on the multifaceted landscape of AI, exploring its paradigms, interplay with human agents, and the overarching prospects and challenges that mark the path forward.

2 A Short History of AI

In contrary to what is commonly accepted, AI might not started entirely in the mids of the 20th century. Rather, it is acceptable to say that it flourish there, but the seeds where planted centuries before that. Indeed, the ever-evolving field of AI has a rich and fascinating history that dates back to ancient times [42]. In fact, AI has captured the imagination of researchers, scientists, philosophers, and enthusiasts for centuries. From its ancient philosophical roots to the modern era of its subfields of ML, the history of AI is a captivating tale of human ingenuity and pursuit of intelligent machines. In this section, we embark on a journey through time, exploring the ancestral events and key milestones that have shaped the development of AI to its present stage.

The concept of artificial beings endowed with intelligence or consciousness can be traced back to ancient myths, stories, and rumors. In fact, the quest for creating artificial beings with intelligence begun in antiquity, in the ancient civilizations such as Greece and Egypt, which were fascinated by the idea of automata and mythical creatures that exhibited lifelike behaviors.

In Greek mythology, there were tales of automatons created by master craftsmen, such as Talos, a giant bronze automaton, first mentioned around 700 B.C. by Hesiod, or Pandora, an artificial woman created by Hephaestus that was given as a gift to Epimetheus and her curiosity led her to open a box, releasing all the evils into the

world. Thus, long before the modern AI, these early narratives laid the foundation for the idea of artificial beings with human-like capabilities [42]. Long before the current discussion on trustworthy and responsible AI, this myth explores the consequences of creating an artificial being with human-like qualities. Another example comes also from Hephaestus' automated servants, which according to Homer's accounts, created a set of automated servants who looked like women but were made of gold. These artificial creatures were given the gods' knowledge and can be considered an ancient mythical version of AI. Furthermore, philosophers such as Aristotle, in his work "On the Soul", explored the concept of the "nous", a rational intellect that distinguished humans from other beings. These early philosophical musings laid the groundwork for future contemplation on the nature of intelligence. The seeds for AI were created and just need to be planted. And this was done later by other philosophers who attempted to describe human thinking as the mechanical manipulation of symbols. Philosophers like René Descartes [18] and Gottfried Leibniz [38] explored the idea of a universal language of thought and the possibility of creating thinking machines. Regarding Descartes, his philosophical ideas on the mind-body problem and the nature of consciousness have influenced the study of AI. His concept of dualism, which separates the mind and body, has implications for understanding the relationship between human intelligence and machine intelligence. These philosophical inquiries set the stage for future developments in AI.

But the work of philosophers, shaping the understanding and development of AI, is not confined to the past. It is fair to say that they currently continue to explore the philosophical implications of intelligent machines. For instance, more recently, philosophers such John Searle [58] or David Deutsch [19], among others, have contributed also to the philosophical foundations of AI, exploring questions about intelligence, consciousness, ethics, and the nature of mind, and thus influencing ongoing research and discussions in the field. More precisely, Searle's thought experiment known as the Chinese Room argument argues that a computer program, no matter how sophisticated, cannot truly understand or possess consciousness. This argument challenges the idea of strong AI (see Section 3.1.1), which posits that machines can have genuine mental states. On the other hand, Deutsch, a physicist and philosopher, has argued that the development of artificial general intelligence is a matter of philosophy rather than just computer science or neurophysiology. He believes that philosophical progress is essential for the integration of artificial general intelligence into society.

In order to better flourish, the seeds of AI as a science and engineering field planted by these philosophers and storytellers needed a substrate that appeared only in the 1930s, 1940s, and 1950s, curiously and precisely in one of the hardest period faced by humanity: the World War II, its prelude and postwar aftermath. It is at this turbulent period of the 20th century that the true birth of AI is placed, when several pioneering events marked the inception of the field. Early pioneers, such as Alan Turing, made groundbreaking contributions to the field by proposing the concept of a universal machine capable of simulating any other machine. The idea was published in a paper in 1936 titled "On Computable Numbers, with an application to the Entscheidungsproblem" [69]. Considered as a foundational work in computer

science and often cited as the basis for the development of modern computers, in this paper, Turing presented a theoretical machine that could solve any problem that could be described by simple instructions encoded on a paper tape. This machine, now known as a Turing machine, was a theoretical model of a general-purpose computer that could perform any computation that could be described by an algorithm. The paper also introduced the concept of computability and the idea that some problems are not computable. Turing's work on computability and the Turing machine laid the foundation for the development of modern computers and AI.

Later on, in 1950, Alan Turing's seminal paper "Computing Machinery and Intelligence" [70] introduced the idea of a test for machine intelligence, now famously known as the Turing Test. This is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. More precisely, in the Turing Test, a human judge engages in a natural language conversation with both a human and a machine, without knowing which is which. If the judge cannot reliably distinguish which is the human and which is the machine, the machine is said to have passed the Turing Test. This is essentially a test of the machine's ability to mimic human-like conversational responses. Turing proposed that if a machine could exhibit human-like behavior in conversation, it could be considered intelligent. This is to some extent related with Searle's Chinese Room scenario. The relationship between these concepts lies in the broader discussion of AI, understanding, and consciousness. Searle's Chinese Room argument is a philosophical challenge to the idea that mere symbol manipulation (as in a computer program) can equate to genuine understanding, which has implications for the Turing Test in the context of assessing the capabilities of AI. Searle's argument suggests that passing the Turing Test does not necessarily imply true understanding or consciousness in a machine. Overall, while Searle's Chinese Room argument and the Turing Test both explore the limits of AI and machine understanding, they approach the topic from different angles, with the Chinese Room argument critiquing the idea of understanding in symbol manipulation, and the Turing Test focused on evaluating conversational abilities.

For all these reasons, in the constellation of stars formed by those philosophers and storytellers who have contributed to AI and all those scientists that follow them, the name of Alan Turing sparks. Turing is often considered the father of AI, although the strong statement that AI has a single father might be object of critics, as the next lines attempts to show. Anyway, his groundbreaking work on computability and the concept of a universal machine laid the foundation for AI research.

Warren McCulloch and Walter Pitts, another two AI stars, introduced the concept of artificial neurons in their paper titled "A Logical Calculus of Ideas Immanent in Nervous Activity", published in 1943 [43]. This paper proposed the first mathematical model of a neural network, often referred to as the McCulloch-Pitts neuron. Their work provided a way to describe brain functions in abstract terms and demonstrated that simple elements connected in a neural network could have immense computational power. Although their paper initially received little attention, its ideas were later applied by influential figures such as John von Neumann and Norbert Wiener [48]. The McCulloch-Pitts neuron served as a precursor to the development of neural networks and ML tools that we have today.

The field of AI research was officially founded at a workshop held on the campus of Dartmouth College in the summer of 1956: the Dartmouth Summer Research Project on Artificial Intelligence. At this historic meeting, John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon coined the term "artificial intelligence" and officially established AI as a distinct field of study. This workshop brought together leading researchers, including, and in addition the aforementioned, Allen Newell, and Herbert Simon, who would become influential figures in AI research for decades to come. The participants at the Dartmouth workshop were optimistic about the future of AI, with many predicting the creation of a machine as intelligent as a human within a generation. The conference set forth ambitious goals, envisioning machines capable of emulating human intelligence and solving complex problems.

But let us go back a little and look briefly at the history of computers that enabled the operationalization and the transition from those theoretical flourishing ideas that were in the origins of AI to their practical implementation. Although the invention of the first computer is a topic of debate, as there are different classifications of computers, and that the Antikythera mechanism, a hand-powered mechanical device, dating back to 200 BC - 70 BC, is believed to be the first device resembling modern machines that we know about, the first mechanical computer is attributed to Charles Babbage [68]. The Analytical Machine, as it is known, was invented in 1822, but it probably does not resemble what most would consider a computer today. Much more closer to modern computers, the first programmable digital computer appeared in the 1940s and marked a significant milestone in the history of AI. This machine, based on mathematical reasoning, provided a platform for exploring the principles of AI. The seed of AI planted by the ancient philosophers and storytellers had now an important substrate to flourish.

During the very beginning of electronic computers, coincidence or may be not, this period marked the birth of symbolic AI, with early attempts to program intelligent behavior using symbolic rules and logic. In the following years, AI research focused on developing practical applications. The first shown AI system is "Theseus", Claude Shannon's remote-controlled mouse that was able to find its way out of a labyrinth and could remember its course [60].

The history of AI has seen periods of both excitement and disappointment [50, 56]. As the initial optimism surrounding AI grew, the field faced a period of skepticism and reduced funding, often referred to as the "AI winter". More precisely, from the 1970s to the 1980s, progress in AI research slowed, and the ambitious goals set by the Dartmouth Summer Research Project on Artificial Intelligence seemed out of reach. Symbolic AI struggled to handle uncertainty and real-world complexity, contributing to the wane in interest and support.

In the 1980s, AI research shifted towards expert systems and knowledge-based AI. Researchers focused on capturing the expertise of human specialists in rule-based systems. Expert systems showed promise in specific domains like medicine and engineering, but their limitations in dealing with uncertain or incomplete information hindered their widespread adoption.

The 1990s marked a resurgence of interest in AI, often referred to as the AI renaissance. Advances in ML and statistical methods, coupled with increased com-

putational power, revitalized the field. Researchers explored probabilistic reasoning, Bayesian Networks, and various learning algorithms to create more robust and adaptive AI systems.

The early 21st century witnessed a shift towards ML as a dominant paradigm in AI. Breakthroughs in deep learning, driven by advancements in neural networks and computational resources, revolutionized computer vision, natural language processing, and other AI applications. The availability of vast datasets and the rise of big data further propelled the success of ML methods.

In recent years, AI has emerged as a transformative force across various domains, deeply embedding itself into our daily lives. This technological advancement has led to the creation of a multitude of AI applications, which have revolutionized industries ranging from healthcare and finance to entertainment and transportation [74, 4, 11]. As AI technologies continue to evolve, they offer unprecedented opportunities for innovation and problem-solving across the spectrum of human endeavors. For instance, AI-driven language translation tools like Google Translate have broken down communication barriers, enabling real-time translation between languages. This has profound implications for global business, diplomacy, and education. Additionally, AI-powered recommendation systems employed by streaming platforms like Netflix and music services like Spotify analyze user preferences to deliver tailored content suggestions, enhancing user experiences and engagement.

In healthcare, AI is being leveraged for early disease detection, personalized treatment plans, and drug discovery. For instance, IBM's Watson for Oncology analyzes vast amounts of medical literature and patient data to assist oncologists in making informed treatment decisions. Similarly, AI-powered diagnostic tools like skin cancer detection apps use image recognition algorithms to identify potential malignancies from photographs.

In the field of autonomous vehicles, AI plays a pivotal role in developing self-driving cars. Companies like Tesla, General Motors, and Waymo utilize advanced computer vision and ML algorithms to enable vehicles to navigate and make real-time decisions on the road. Such advancements have the potential to significantly reduce accidents and improve traffic flow.

Moreover, AI's influence extends to environmental conservation. AI algorithms are applied to analyze satellite imagery and monitor deforestation, wildlife habitats, and climate change. These insights help scientists and policymakers make informed decisions for preserving the planet's biodiversity.

Milestones like IBM's Deep Blue [9] and Google's AlphaGo [64] have exemplified the capabilities of AI systems in ways that were previously thought to be purely within the realm of human expertise. Deep Blue's victory over Garry Kasparov in chess marked a watershed moment, showcasing that AI could outperform even the world's greatest human chess player. Similarly, AlphaGo's triumph over the world champion Go player, Lee Sedol, demonstrated AI's ability to master the complex and intuitive game of Go. These achievements not only highlighted the extraordinary computational power of AI but also challenged our perceptions of what is achievable in the realm of strategic thinking and problem-solving.

However, alongside its numerous benefits, AI also presents ethical and social challenges. The concern over AI's impact on employment due to automation, issues of algorithmic bias, and questions about privacy and security have become focal points of discussion and debate.

In conclusion, AI's rapid evolution has ushered in a new era of possibilities and challenges. Its applications span across industries and fields, reshaping the way we live, work, and interact. As AI continues to advance, its relevance in science, humanities, and beyond will only continue to grow, emphasizing the importance of responsible and ethical AI development to harness its potential for the betterment of society. Its history has been written as a rich tapestry of ideas, contributions, innovations, and a testament to human curiosity and ingenuity that span centuries. From ancient philosophical ponderings to the modern era of ML and neural networks, AI has come a long way. The journey is far from over, as researchers continue to push the boundaries of what is possible in creating intelligent machines. While challenges and ethical considerations remain, AI continues to shape our world and holds immense potential for the future. As we delve deeper into the field, it is essential to learn from the past and build upon the diverse array of methodologies and paradigms that define the history of AI.

3 Artificial Intelligence Paradigms

The field of AI is a vast and dynamic landscape, encompassing various paradigms and methodologies that drive the development of intelligent systems. From symbolic reasoning to ML, AI paradigms represent distinct approaches to solving complex problems and simulating intelligence, mostly human-like but not necessarily. In this section, we explore the key AI paradigms, some of them already placed in the timeline of the history of AI described in the previous section. By AI paradigms we mean the different concepts of intelligence and methodologies used in developing intelligent computer systems, or, in other words, they represent fundamental assumptions, perspectives, and models of intelligence that guide research and system development. Thus, each paradigm refers to a high-level approach, philosophy, or framework for developing AI systems and capabilities.

3.1 Key Paradigms of AI

The field of AI is diverse and constantly evolving, with researchers exploring new methodologies and combining different techniques to advance the capabilities of AI systems. Different sets of paradigms has been proposed in AI by different lines of thought. This diversity reflects the interdisciplinary nature of AI, which draws on ideas and techniques from computer science, mathematics, philosophy, psychology,

neuroscience, and other fields. In line, different categorizations of AI paradigms has been proposed.

3.1.1 Russell and Norvig's four types of AI

Russell and Norvig's [56] four types of AI (thinking humanly, thinking rationally, acting humanly, acting rationally) are not only attempts to define AI, but also they can be considered as key paradigms of AI, defined based on the goals behind the type of AI, in turn defined by two dimensions: one dimension is whether the goal is to match human performance, or, instead, ideal rationality; the other dimension is whether the goal is to build systems that reason/think, or rather systems that act or behave. These four types of AI rely on different concepts of intelligence and employ different methodologies to achieve it, and therefore are key paradigms of AI.

The Thinking Humanly paradigm focuses on creating AI systems that can mimic human thinking processes. It aims to understand and replicate human cognition, including perception, reasoning, and learning. The goal is to develop AI agents that can think like humans.

The Thinking Rationally paradigm emphasizes the use of logical reasoning and symbolic representations. It involves modeling thinking as a logical process, where conclusions are drawn based on logical rules and deductions. The focus is on creating AI agents that can think logically and make rational decisions.

The Acting Humanly paradigm is concerned with creating AI agents that can mimic human behavior. It involves understanding and replicating human actions, gestures, and responses. The goal is to develop AI agents that can interact with humans in a natural and human-like manner.

The Acting Rationally paradigm focuses on creating AI agents that can act in a way that maximizes their chances of achieving their goals. It involves making decisions based on rationality and optimizing outcomes. The focus is on creating AI agents that can consistently manifest rational decisions and execute suitable actions accordingly.

These four paradigms of AI provide different perspectives and approaches to achieving intelligent behavior. They highlight the diverse ways in which AI can be developed and applied.

A fundamental concept in the field of AI particularly related with these paradigms is that of the dichotomy of strong and weak AI, or general and narrow AI [58, 55, 17, 21]. It refers to different levels of AI capabilities and consciousness. It is a philosophical and conceptual distinction that helps us understand the nature of AI systems and their relationship to human intelligence. Weak AI, also known as narrow AI, refers mostly to AI systems designed for specific tasks or problem-solving. These AI systems simulate human intelligence in a limited domain but do not possess consciousness or general intelligence. They operate based on algorithms and data, lacking true understanding. Weak AI systems excel in narrow, well-defined domains but cannot transfer their knowledge or skills to other areas. Examples include virtual personal assistants like Siri, image recognition algorithms, and recommendation

systems. Weak AI aims to automate tasks efficiently and accurately within predefined boundaries, without consciousness.

Strong AI, also known as artificial general intelligence or true AI, represents AI systems with human-level intelligence and consciousness. These systems possess general cognitive abilities such as reasoning, problem-solving, learning, and understanding, across a wide range of domains. Such systems demonstrate common sense, make judgments, solve problems flexibly across domains, learn new skills like humans, exhibit adaptability and can transfer that knowledge and those skills from one domain to another, demonstrating versatility. It may have a form of consciousness, self-awareness, and the potential for subjective experiences and emotions. The goal of strong AI is to create machines that emulate human-like general intelligence, potentially revolutionizing technology and society. However, achieving strong AI remains a theoretical aspiration, with significant ethical considerations regarding the rights of conscious AI entities, existential risks, and societal impacts.

3.1.2 Symbolic, Connectionist, and Statistical AI

Another very common division of AI, based on different features, has been that of splitting AI into symbolic and connectionist AI [49, 35, 3].

Symbolic AI, also known as classical AI or rule-based AI, is rooted in symbolic logic and formal reasoning. This paradigm represents knowledge using symbols, rules, and logical expressions. Early AI systems were built using rule-based expert systems, which encoded human expertise as if-then rules to make decisions and solve problems. The Prolog programming was a powerful tool for developing such systems.

Connectionist AI, inspired by the structure and function of the brain's neural networks, represents a paradigm centered around artificial neural networks. These networks consist of interconnected artificial neurons that process information and learn from data. Connectionist AI has seen significant advances in areas such as pattern recognition, computer vision, natural language processing, and reinforcement learning.

In spite of this classical and well accepted binary division, another paradigm has also gained consensus [45, 56]. This is the statistical AI, often referred to as ML, and is focused on using statistical techniques to enable machines to learn from data and make predictions or decisions. This paradigm includes various learning methods such as supervised learning, unsupervised learning, and reinforcement learning. Statistical AI has become the cornerstone of modern AI applications, including recommendation systems, image recognition, and language translation.

It is important to note that statistical AI, represented by ML techniques, can be applied within both symbolic AI and neural computing paradigms. In fact, ML algorithms can be used to learn rules and logical relationships (symbolic AI) as well as to train and optimize neural networks (connectionist AI). Furthermore, in contemporary AI research and practice, the boundaries between these paradigms are often blurred, and hybrid approaches that combine symbolic reasoning with statistical

learning or neural networks are increasingly common. This integration of paradigms showcases the flexibility and adaptability of AI techniques to address a wide range of problems and challenges. Concluding, while symbolic AI, neural computing, and statistical AI can be seen as distinct paradigms, they are interconnected and often coexist within modern AI systems, with statistical AI techniques (ML) playing a significant role in advancing both symbolic reasoning and neural network models.

It is worth of notice that other similar divisions of AI has also been done, namely that of separating symbolic AI into the logistic-based and the non-logistic based symbolic AI. From this point of view, while the former corresponds to pure logic-based AI, the latter is based on the Bayesian formalism [53, 16] and thus it refers to the probabilistic/statistical AI. In this scenario of logistic symbolic AI and non-logistic symbolic AI, there is still room for a third paradigm, which is as previously the connectionist/neurocomputing paradigm. From this point of view, the statistical AI paradigm of the previous triad division of AI falls in the symbolic paradigm. Various arguments have been expressed to weight positively in favor of this position as follows. Unlike logicist symbolic reasoning, where rules are deterministic and binary, non-logicist symbolic reasoning introduces probabilities to quantify uncertainty. These probabilities reflect the degree of belief or confidence in a particular hypothesis or decision. It allows AI systems to reason and make decisions in situations where the information is uncertain or incomplete. It enables machines to assess the likelihood of different outcomes and choose the most probable course of action based on available evidence, offering a more versatile and robust approach to intelligent decision-making. Probabilistic graphical models, such as Bayesian networks, are commonly used to represent uncertain relationships between variables.

In summary, the subdivision of symbolic AI into logicist and non-logicist reasoning reflects the evolving nature of AI research, particularly in handling uncertainty. While logicist symbolic reasoning relies on deterministic rules and formal logic, non-logicist symbolic reasoning embraces probabilistic methods to reason in the face of uncertainty. Both paradigms (or sub-paradigms of the symbolic paradigm) contribute valuable perspectives to the broader field of AI, highlighting the importance of flexibility and adaptability in building intelligent systems that can thrive in real-world environments. In this perspective, Bayesian networks are considered symbolic because they use discrete symbols to represent variables and relationships in the form of nodes and edges in a graph. The symbolic nature of Bayesian networks is evident in the following aspects. In a Bayesian network, each node represents a variable that can take on a finite set of discrete values. For example, in a medical diagnosis system, nodes might represent symptoms (e.g., fever, cough) and medical conditions (e.g., flu, pneumonia). Each node has a discrete set of possible states, corresponding to different values the variable can take. In addition, the connections between nodes in a Bayesian network are represented by directed edges. These edges indicate probabilistic dependencies between variables, showing the causal or conditional relationships among them. The directed nature of the edges implies that the network encodes a particular directionality or causal influence from parent nodes to child nodes. Each node in a Bayesian network has an associated conditional probability table. The conditional probability table specifies the conditional probabilities

of the node's possible states given the states of its parent nodes, this way encoding the probabilistic relationships between variables and provide a way to reason and update beliefs in the face of new evidence. Bayesian networks are widely used for knowledge representation and reasoning under uncertainty. They provide a structured and interpretable way to model uncertain knowledge in a domain.

In the triad division of AI into symbolic, connectionist and statistical, while Bayesian networks represent a form of symbolic AI, it is essential to note that they also embrace aspects of statistical AI. They combine symbolic representations with probability theory to enable reasoning and decision-making in uncertain and complex environments. As a result, in this triad division, Bayesian networks serve as a bridge between symbolic AI and statistical AI paradigms, showcasing the integrative nature of AI methodologies in addressing real-world challenges.

3.1.3 Pedro Domingo's five tribes of AI

In spite of the aforementioned categorizations of AI described in the previous section, one, more classical, with two paradigms (Symbolic and Connectionist AI) and another with three (Symbolic, Connectionist and Statistical AI/ML), Pedro Domingos [21] has recently proposed a five paradigm scenario in the context of ML. To those five paradigms he called tribes of AI. Each one of these tribes offer a different definition for intelligence and a different methodology to understand and create intelligent machines. The paradigms are:

- Symbolic AI: This paradigm represents the use of rules and logic to simulate human reasoning. It is based on representing knowledge using symbols and applying logical operations to draw inferences and make decisions.
- Connectionist AI (Neural Networks): Connectionist AI focuses on artificial neural networks, which are inspired by the structure and functioning of the brain's neural networks. These networks consist of interconnected artificial neurons that learn from data and adapt their connections to improve performance.
- Evolutionary AI: Evolutionary AI draws inspiration from biological evolution to optimize solutions to complex problems. It involves using genetic algorithms and genetic programming to evolve populations of candidate solutions through processes like selection, crossover, and mutation.
- Bayesian AI: This paradigm is centered around Bayesian networks, which, as described above, are probabilistic graphical models representing uncertain relationships between variables. Bayesian AI involves reasoning under uncertainty and updating beliefs based on new evidence.
- Analogical AI: Analogical AI focuses on reasoning by analogy, where knowledge from past experiences is used to solve new problems by finding similarities and patterns in different domains.

Domingos has proposed that these paradigms can be combined to create a "master algorithm" capable of learning nearly anything. He has also emphasized the impor-

tance of a multidisciplinary approach to ML, drawing on insights from fields such as philosophy, biology, neuroscience, statistics, and psychology.

3.2 Emergent Paradigms

Either the classical triad of paradigms or Russell and Norvig's four paradigms, or even the five paradigms mentioned by Domingos, provide a useful framework for understanding the dominant paradigms and approaches in AI history. However, AI is a rapidly evolving field, and new paradigms or sub-paradigms can emerge as researchers explore novel methodologies and techniques. Thus, the following emerging or alternative paradigms in AI may deserve attention in the future, or already in the present, such explainable, ethical, trustworthy and responsible AI [20, 40, 32]. Although closely related, these concepts present differences, as explained as follows.

Explainable AI (XAI) focuses on developing AI systems that can provide understandable and interpretable explanations for their decisions and actions. The goal is to make AI more transparent, trustworthy, and accountable. XAI research aims to bridge the gap between the black-box nature of certain AI models and the need for human comprehensibility.

Ethical AI focuses on developing AI systems that are aligned with ethical principles and societal values. This tribe addresses concerns related to biases, fairness, privacy, accountability, and the social impact of AI. Ethical AI research aims to ensure that AI technologies are developed and deployed responsibly and consider the ethical implications of their use.

There are other contemporaneous paradigms that promise to revolutionize computing and consequently AI. That is the case of Quantum AI [72, 5, 22, 57], which explores the intersection of AI and quantum computing. This tribe investigates how quantum computing can enhance AI algorithms and solve complex problems more efficiently. Quantum AI aims to leverage the unique properties of quantum systems, such as superposition and entanglement, to develop more powerful AI models and algorithms.

3.3 Summary

In AI, there is a growing interest in heterogeneous approaches that combine various formalisms. The field has embraced a fundamental triad of paradigms: logicist, probabilistic/Bayesian, and neurocomputational techniques. The resurgence of Bayesian and neurocomputational methods has further strengthened their position alongside the original top-down logicist paradigm. Now, AI systems, such as Watson's DeepQA [24] and Google DeepMind's AlphaGo [63], effectively leverage multiple paradigms. Although deep theoretical integration of these paradigms remains a future possibility,

researchers are actively striving towards this goal. For instance, efforts are underway to demonstrate how neurocomputational processes can give rise to symbolic cognition. Additionally, cognitive architectures like Soar [36] and PolyScheme [10] are contributing to the integration of diverse AI fields. The Companions project [26] is one such initiative, working to build human-level AI systems capable of collaborative problem-solving and long-term interaction with humans.

AI is evolving and new paradigms are appearing. Classic or modern, each paradigm offers a unique perspective on the concept of intelligence and the methods used to create intelligent computer systems, but when blended they may merge the best of them.

4 Artificial, Intelligent, Autonomous Agents

Russell and Norvig [56] highlight intelligent agents and the percept-action mapping function as a common framework for studying and implementing AI across different techniques from symbolic to statistical approaches. Thus, agent formulation provides a unifying lens for different AI techniques, representing a problem as perceiving and acting rationally to achieve goals. From this point of view, more than a paradigm, the intelligent agent concept is a common denominator to all paradigms. Different AI techniques, originating from different paradigms, can be used to implement the agent mapping function, including symbolic reasoning, machine learning, neural networks, etc. Furthermore, these different paradigms can be used in different components/modules of an AI agent architecture.

The next sections will provide a more precise definition of intelligent agents, the characterization of the tasks they are endowed to do and the environments in which they operate (or, more strongly,"live"), the different categories of their architectures, from simplistic to more complex ones, including, for their relevance to AI, the machine learning one and the architecture that endow them with human mental qualities such as beliefs, desires, intentions, and their utilization to make-decisions in Bratman's practical reasoning model [8]. Finally, we will describe how these agents may be joined together and provided with additional social capabilities such as coordination, collaboration, cooperation, negotiation, and competition, forming a multi-agent system.

4.1 Definition

Artificial intelligent autonomous agents [56, 73] are computational entities or systems designed to interact with their environment, perceive information through sensors, process that information, and autonomously, i.e., without direct human intervention or from any other external agents, take actions to achieve specific goals

or objectives. These agents possess the ability to make decisions, learn from their experiences, and adapt their behavior based on changes in their environment.

There is no full consensus on the list of features that characterise artificial autonomous agents, but the following receive some consensus in the literature [73]:

- Autonomy: Agents have a degree of independence and are capable of acting without constant human intervention or from other agents. They can perceive the environment, make decisions, and execute actions on their own.
- Reactivity: Agents are responsive to changes in the environment. They can perceive and respond to incoming sensory data or external stimuli in real-time.
- Proactiveness: Agents are not solely reactive but can also exhibit proactive behavior. They can take the initiative to achieve their goals by executing appropriate actions, rather than simply reacting to the environment.
- Goal-Oriented: Agents have specific objectives or goals that they seek to achieve. Their actions are directed towards optimizing their performance measure with respect to these goals.
- Communication: Agents can communicate with other agents or entities in the environment. This communication can be through direct interaction or through the exchange of messages.
- Learning and Adaptation: Many agents have the ability to learn from past experiences or interactions with the environment. They can adapt their behavior over time to improve their performance.
- Social Ability: In multi-agent systems, agents may have social abilities, allowing them to interact, cooperate, and negotiate with other agents to achieve common or conflicting goals.

An important concept that gains prominence when talking about artificial agents is that of rationality. In the context of autonomous artificial intelligent agents, it refers to the property of making decisions that maximize the expected utility based on the available knowledge and goals of the agent [56]. It should be keep clear that rationality is not about being omniscient or having complete knowledge of the world but rather about making the best possible decision given the information and resources available to the agent at any given moment. An agent is considered rational if its actions are in line with its goals, and it selects actions that are likely to lead to the most favorable outcomes or higher expected utility. The concept of rationality is closely related to the idea of acting intelligently. An agent can be viewed as intelligent if it behaves rationally, i.e., it chooses actions that are coherent with its objectives and the current state of the environment. This perspective allows for the assessment of an agent's intelligence based on the outcomes of its actions and its ability to adapt and improve over time.

4.2 Tasks and Environments

Rational agents "live" in an environment and do some tasks. So, in order to design an agent, one has to provide the characterization of the task environment, i.e., the problem that the agent has to solve. In order to do that the following PEAS (Performance, Environment, Actuators, and Sensors) framework is used. It provides a systematic way of characterizing an agent's behavior, its interaction with the environment, and the means by which it can affect and perceive the world. PEAS comprises the following dimensions:

- Environment: The environment is the external context in which the agent operates. It includes everything that the agent interacts with and influences, and it may vary in complexity, dynamism, and unpredictability. The environment provides inputs to the agent through sensors and receives outputs from the agent through actuators. The nature of the environment heavily influences the agent's decision-making and problem-solving capabilities.
- Performance: Performance refers to the measure of success or the criteria used to evaluate how well an agent is achieving its objectives. It represents what the designer wants the agent to accomplish in a given environment. For example, in a chess-playing agent, the performance measure might be the number of games won against opponents or the average number of moves required to win a game.
- Actuators: Actuators are the means by which the agent can effect changes in the environment. They are the channels through which the agent takes action or performs actions in response to its internal decision-making processes. Actuators can be as simple as motors or more complex as robotic arms or natural language generators, depending on the agent's capabilities and the tasks it needs to perform.
- Sensors: Sensors are the components that allow the agent to perceive and gather information about the environment. They are responsible for collecting data from the environment and transforming it into a format that the agent can process and use in its decision-making. Sensors can include cameras, microphones, temperature sensors, and other types of detectors, depending on the agent's sensing capabilities.

By defining the PEAS components, designers and developers can gain a clear understanding of an agent's goals, its interaction with the environment, and the mechanisms it uses to achieve its objectives. This framework serves as a basis for discussing and comparing different agent designs and helps in evaluating how well an agent performs in a given context. The PEAS framework is widely used in the field of AI to model and analyze various intelligent systems and their behaviors in different environments.

4.3 Properties of environments

After defining the task environment, it is advisable to analyse the environment itself from different dimensions [56]. The following ones are usually considered:

- Observability: this dimension involves determining whether the agent's sensors provide access to the complete state of the environment. An environment is considered fully observable if the agent can directly observe all relevant aspects of the environment that are necessary for decision-making. If the environment is only partially observable, the agent may have to maintain internal models or beliefs about the unobservable parts.
- Single vs. Multi-agent: this dimension identifies whether the environment contains just one agent (single-agent environment) or multiple agents (multi-agent environment) that can interact and compete with each other. In multi-agent environments, the behavior of other agents becomes a crucial factor in the agent's decision-making.
- Deterministic vs. Stochastic: the focus is on whether the environment is deterministic or stochastic. In a deterministic environment, the next state is uniquely determined by the current state and the agent's actions. In contrast, a stochastic environment has an element of randomness or uncertainty in its transitions.
- Episodic vs. Sequential: the focus is on whether the agent's performance is measured by individual episodes (episodic environment) or over a series of actions and decisions (sequential environment). In episodic environments, the agent's actions have no influence on future episodes, while in sequential environments, actions can have long-term consequences.
- Static vs. Dynamic: this dimension examines whether the environment changes while the agent is deliberating or making decisions. A static environment remains constant throughout the agent's decision-making process, whereas a dynamic environment can change unpredictably.
- Discrete vs. Continuous: this dimension classifies the state, action, and time dimensions of the environment as discrete or continuous. In a discrete environment, there are a finite number of states, actions, and time steps. In contrast, a continuous environment has an infinite number of possibilities for states, actions, or time.
- Known vs. Unknown: this dimension looks at whether the agent has complete knowledge about the environment's dynamics and rules (known environment) or whether some aspects are unknown to the agent (unknown environment).

By characterizing an environment using these dimensions, AI designers can gain insights into the complexity and challenges an agent might face while operating in that environment. This characterization helps guide the design and selection of appropriate AI algorithms and strategies to ensure effective and efficient decision-making and action-taking by the agent.

4.4 Architecture and Taxonomy of Agents

Agents can be classified based on their structure or architecture and are often categorized into different types based on their decision-making processes and capabilities. One of the main classification criteria distinguishes between deliberative and reactive agents [73]. Deliberative agents are characterized by their ability to think and plan before taking actions. They employ internal models of the environment to reason about different possible sequences of actions and their potential outcomes. These agents typically have a long-term perspective and consider the consequences of their decisions. Deliberative agents use sophisticated algorithms, such as search and planning, to reach informed decisions.

On the other hand, reactive agents act purely in response to the current percept, without maintaining an internal model of the environment. These agents react quickly to their current sensory input and rely on pre-defined rules or mappings from perceptions to actions. Reactive agents are particularly well-suited for real-time tasks and situations where immediate action is crucial. This is the case of some robots designed to perform tasks such as obstacle avoidance, path planning, and object recognition. Autonomous vehicles are reactive in specific situations when performing tasks such as lane keeping, collision avoidance, and traffic signal recognition.

There is a third class, called hybrid agents, that combine the strengths of both deliberative and reactive approaches. They maintain a balance between thinking and acting, employing a combination of planning and reactive strategies. Hybrid agents adapt their decision-making process based on the complexity of the task and the available computational resources.

Quite related with this classification method, there is another that assumes a correspondence between simple reflexive agents and reactive agents, and that splits the deliberative agents into various ones. The resulting taxonomy of agents is described as follows [56]:

- Reflex agents are a type of reactive agent that maps directly from percepts to actions based on predefined rules or condition-action pairs. They do not possess memory or internal state, making their behavior solely determined by the current percept. While simple, reflex agents are effective for tasks with fixed environmental conditions and limited complexity. If-then-else rules govern their decision-making process.
- Model-based reflex agents extend the capabilities of reflex agents by incorporating an internal model of the environment. This model allows them to reason about the outcomes of different actions and make more informed decisions. Despite this added complexity, they still react quickly based on the current percept.
- Goal-based agents have explicit goals that guide their decision-making process. They use their internal model of the environment to plan actions that will lead them towards achieving their objectives. Goal-based agents are more flexible than reflex agents and can adapt to different environmental conditions.
- Utility-based agents optimize their actions based on a utility function that assigns a numeric value to different outcomes. By evaluating the desirability of various

states, utility-based agents can select actions that maximize the expected outcome, even in uncertain environments.
- Learning agents improve their performance over time through experience. They can adapt their behavior based on feedback from the environment, allowing them to learn from both successes and failures. Learning agents employ various ML techniques to update their decision-making strategies. the next section presents more details on these techniques.

4.5 Machine Learning Agents

Specifically regarding the latter category of agents listed in the previous section, there are numerous criteria for categorizing ML techniques. Active versus passive learning, online versus offline learning, instance-based versus model-based learning, deductive versus inductive learning, symbolic versus subsymbolic/connectionist learning, are just examples [56, 45]. However, probably the most common is splitting them into the following three main categories/branches according to their learning paradigms and data interactions: supervised learning, unsupervised learning, and reinforcement learning. Each of these categories represents a distinct approach to ML, addressing different types of problems and learning scenarios. Many real-world applications involve a combination of these approaches to achieve desired outcomes. As ML continues to evolve, new techniques and hybrid approaches are being developed to tackle complex and diverse challenges.

4.5.1 Supervised Learning

Supervised learning is the most common and well-known category of ML. In supervised learning, the algorithm learns from labeled training data, where each input example is associated with a corresponding target or label. The goal of the algorithm is to learn a mapping from inputs to outputs, based on the patterns present in the labeled training data. The algorithm generalizes from the training data to make predictions or classifications on new, unseen data. The learning process involves a clear distinction between input features and output labels. The algorithm is trained to minimize the discrepancy between predicted outputs and actual labels. Common applications encompass image classification, email spam detection, medical diagnosis, emotion detection, and various others.

4.5.2 Unsupervised Learning

Unsupervised learning involves working with unlabeled data, where the algorithm's goal is to uncover underlying patterns, structures, or relationships within the data. Unlike supervised learning, there are no target labels to guide the learning process.

Instead, the algorithm aims to discover inherent groupings, clusters, or dimensions in the data. There are no predefined target labels, and the algorithm focuses on understanding data relationships. Common techniques include clustering, dimensionality reduction, and anomaly detection. Unsupervised learning is often used for exploratory data analysis, customer segmentation, and data compression.

4.5.3 Reinforcement Learning

Reinforcement learning [67] is a paradigm inspired by behavioral psychology, where an agent learns how to interact with an environment to maximize cumulative rewards over time. The agent takes actions in the environment, and the environment provides feedback in the form of rewards or penalties based on the agent's actions. The agent learns to make decisions that lead to the highest cumulative reward. The agent learns by trial and error, exploring different actions to learn optimal policies. Striving for a balance between exploration (trying new actions) and exploitation (choosing actions that yield higher rewards) is key to achieving success in this task. Reinforcement learning is used in scenarios where there is a sequence of decisions and actions to be taken, such as games and robotics [67].

4.6 Beliefs, Desires, and Intentions Architecture

The agent architectures we have examined so far, from simple reflex agents to sophisticated utility-based and learning agents, provide different approaches to designing agents that can perceive environments and select actions. However, most of these architectures lack an explicit representation of the inner cognitive states that drive rational behavior in humans.

This brings us to the influential BDI (Beliefs, Desires, and Intentions) architecture proposed by Bratman, Israel, and Pollack in their 1987 book on practical reasoning [8]. The BDI model aims to capture the psychological traits that characterize purposeful, practical reasoning in people. It includes explicit representations of beliefs, desires, and intentions within the agent.

Beliefs represent the informational state of the agent – what it knows about the world. The beliefs can be incomplete or incorrect. These beliefs are derived from the agent's sensory input and internal state. They can include information about the agent's own state, the state of the environment, and the actions available to the agent. There are beliefs about the past, the present (assumptions) and the future (expectations).

Desires indicate the motivational state in terms of long-term objectives or goals the agent wants to accomplish. They reflect what the agent wants to achieve or accomplish in its environment. Desires can range from simple objectives, such as reaching a specific location, to complex goals involving multiple steps and long-term planning.

Intentions are the deliberative state, i.e., what the agent has chosen to do in the near future. Intentions are driven by the desires and formed based on current beliefs about the world. They are the agent's selected courses of action to fulfill its desires. They represent the agent's commitments to carrying out specific actions based on its beliefs and desires. Intentions guide the agent's behavior and are typically short-term plans of action that can change dynamically as the agent's beliefs and desires are updated.

By modelling these cognitive components, BDI provides a conceptual framework for practical reasoning and means-end analysis in intelligent agents. The architecture has formed the basis for many agent programming languages and implementations in fields like autonomous agents, multi-agent systems, and cognitive robotics. This BDI paradigm aims to support sophisticated, human-like cognitive capabilities in artificial agents.

The BDI architecture emphasizes the role of practical reasoning in decision-making. Agents using the BDI architecture engage in a process of practical reasoning to determine their intentions. This involves reasoning about their beliefs, desires, and available actions to select the most appropriate course of action to achieve their goals.

The BDI architecture also supports plan revision and plan recognition. Agents using this architecture can reconsider their intentions and adjust their plans when new information is received or when the environment changes. Additionally, they can recognize the intentions of other agents by observing their actions and reasoning about their beliefs and desires (see Section 5.3).

One of the key advantages of the BDI architecture is its transparency and human-like reasoning. By structuring the agent's decision-making process around beliefs, desires, and intentions, it becomes easier to understand and validate the agent's behavior, making it more interpretable and explainable.

The BDI architecture has found applications in various fields, including robotics, intelligent agents, and multi-agent systems. It provides a powerful framework for designing autonomous agents capable of reasoning, planning, and adapting to dynamic environments. However, it also has limitations, such as scalability and complexity management, which need to be addressed in more extensive and complex scenarios.

4.7 Multi-Agent Systems and Cooperative Artificial Intelligence

In the realm of AI, multi-agent systems [73, 62, 56, 71, 23, 52, 33], composed of multiple autonomous agents that interact with each other, communicating, competing, negotiating, etc., have emerged as powerful frameworks that permit modeling and simulating complex interactions among autonomous agents. These agents can be computational entities, including softbots, robots, and even humans, or a combination thereof.

Cooperation, collaboration, negotiation, and communication are essential aspects of multi-agent systems and intelligent autonomous agents. These concepts play a

crucial role in enabling agents to work together effectively, achieve common goals, and adapt to dynamic environments.

Negotiation is the process of reaching agreements between autonomous agents with potentially conflicting interests or goals. When agents in a multi-agent system have different objectives, negotiation allows them to find mutually beneficial outcomes and resolve conflicts peacefully. Negotiation involves exchanging proposals, making concessions, and reaching compromises to achieve a consensus. Effective negotiation strategies are crucial for successful interactions between autonomous agents, especially in competitive or cooperative scenarios.

Communication is a fundamental mechanism for agents to exchange information, coordinate actions, and share knowledge. Communication allows agents to update their beliefs, inform others about their intentions, request assistance, and convey their understanding of the environment. Communication can take various forms, such as explicit messages, shared data structures, or implicit signaling through observed actions. Effective communication is essential for maintaining a shared understanding among agents and facilitating effective cooperation and collaboration.

Even though possessing their own individual goals, knowledge, and capabilities, those agents may work together, i.e., they may cooperate to achieve individual or collective objectives (even though there are multi-agent systems in which agents do not cooperate, but rather compete).

In a general sense, cooperation refers to any form of joint effort or working together towards a common purpose or supporting others' goals without a shared objective, i.e., even if the parties involved may not share the same ultimate goal [15]. It encompasses a wide range of interactions, from providing support and assistance to sharing resources or information, without necessarily requiring a tightly coordinated effort. Cooperation plays a vital role in enhancing the problem-solving capacity of multi-agent systems, making them adept at tackling intricate challenges that transcend the capabilities of single entities. It also emphasizes the willingness of agents to assist each other without necessarily having a shared goal, i.e., it may involve working with others to help them achieve their individual goals. Thus, the existence of a shared goal is not mandatory. In cooperation, agents might provide assistance, share resources, or offer complementary expertise to facilitate the accomplishment of individual or mutual objectives.

Even though there are cooperative scenarios in which the agents do help the achievement of other's individual inner goal, there are also many situations in which the agents work towards a common or collective goal, sometimes at the cost of their own individual goals, in case these are different. This is called collaboration. They are benevolent at the point of taking the common or collective goal as their primary goal, which becomes in this way its goal (note that, in cooperative, non-collaborative scenarios, agents never assume the other's goals as their own goal). They work together to achieve that shared objective, i.e., multiple agents or entities come together with a unified purpose or a joint mission. They align their individual interests and efforts towards that common goal, and their success is tied to the successful achievement of that shared objective. The agents collaborate by combining

their knowledge, skills, and resources to collectively work towards a specific outcome that benefits all of them.

Thus, in comparison to collaboration, cooperation is a broader concept that encompasses various forms of joint effort or interaction between agents, regardless of whether they have a shared goal. In cooperative scenarios, agents may provide support, share resources, or assist each other without having to work towards a common objective. Collaboration involves agents with shared goals, working together in a coordinated and interdependent manner to achieve a collective objective. In cooperative scenarios, agents may interact and assist each other, but they may not necessarily be pursuing a common goal. This is a harder situation as, in contrast to the benevolence of collaborative scenarios, in cooperative scenarios the agents might not be benevolent, i.e., they might be self-interested.

In contrast to when they do not share a common goal, collaboration involves a higher degree of coordination and integration among the participants. In a collaborative effort, the entities work jointly on a specific task or project, sharing responsibilities, knowledge, and resources, with the explicit goal of achieving a common outcome. Collaboration often requires more open communication, joint decision-making, and a greater level of interdependence among the participants.

To some extent, collaboration can be seen as a specialized form of cooperation, where the focus is on working together in a more structured and synchronized manner towards a shared objective. While all collaborative efforts involve cooperation, not all cooperative efforts necessarily reach the level of collaboration, as they may involve, for instance, less intensive coordination or shared responsibility.

In conclusion, cooperation is an essential pillar of success for multi-agent systems. It enables autonomous agents to work together efficiently, and leverage collective intelligence for better outcomes, leveraging their diverse knowledge, expertise, and skills to solve complex problems and achieve shared objectives or even individual objectives (in those cases where there is no common objective). They can exchange information, coordinate their actions, and share resources to achieve better outcomes than individual agents acting independently. Collaboration involves communication, coordination, and synergy among agents. It empowers multi-agent systems with the collective intelligence needed to tackle real-world challenges across diverse domains, making them invaluable tools in the realm of AI.

These concepts are integral to the design of multi-agent systems and the development of intelligent autonomous agents. They enable agents to interact in complex and dynamic environments, adapt to changes, and effectively solve problems that require collective efforts. Researchers and developers focus on designing communication protocols, negotiation algorithms, and coordination mechanisms to ensure that agents can work together efficiently and achieve their goals effectively. As multi-agent systems become increasingly prevalent in various applications, understanding and harnessing cooperation, collaboration, negotiation, and communication become crucial for building robust and intelligent autonomous agents.

4.8 Summary

In this section, we have delved into a detailed exploration of intelligent agents, covering several key aspects. We have refined our understanding by defining intelligent agents more precisely, identifying the specific tasks they are designed to perform, and exploring the diverse environments in which they function. Additionally, we have examined the spectrum of agent architectures, ranging from basic structures to complex frameworks like machine learning and those capable of emulating human mental attributes such as beliefs, desires, and intentions, which are pivotal in decision-making processes as outlined in Bratman's practical reasoning model. Furthermore, we have outlined how these individual agents can be integrated to form a cohesive multi-agent system, equipped with advanced social capabilities like coordination, collaboration, cooperation, negotiation, and competition. This comprehensive overview lays the foundation for a deeper dive into the practical applications and implications of these intelligent agents and multi-agent systems in various domains.

5 Cooperation between Human Agents and Artificial Intelligent Agents

In the previous section, we addressed artificial autonomous agents and ended with a special emphasis on cooperation in multi-agent systems. In this section, we delve into one form of cooperative AI, possibly also with collaborative AI contours, which refers to the integration of AI agents and human intelligence to work together towards a common purpose or goal, usually an ultimately goal established by a human, keeping in mind that it does not necessarily imply that both have the same ultimate objective or shared goal. Human-AI cooperation combines the strengths and capabilities of both AI and human agents, leveraging their complementary abilities to achieve superior performance. This particular form of cooperative AI recognizes that while AI systems excel in processing large amounts of data, making fast computations, and identifying patterns, they may lack certain human qualities such as common sense reasoning, creativity, and ethical judgment. On the other hand, humans possess rich contextual knowledge, intuition, and social intelligence, but can be limited by cognitive biases and the capacity to handle massive data sets. By integrating AI and human agents in a cooperative framework, the aim is to achieve outcomes that are more accurate, robust, and aligned with human values. Thus, regarding the prevalence of control (the focus of power of decision-making), the cooperative AI can manifest in various ways such as those addressed in the following subsections.

5.1 Human-in-the-loop vs. Machine in the Loop

There might be two forms of cooperation between AI agents and humans: Human-in-the-Loop (HitL) AI [46], and Human in Command or Machine-in-the-Loop (MitL) [12, 44, 7].

HitL AI [46] can be considered a form of cooperative AI, since it involves continuous interaction and feedback loops between humans and AI systems working together towards shared goals or individual goals. Through this interaction, human input and oversight are integrated into the AI system's pipeline, including data collection and preparation, model training, and especially in the decision-making process. The AI system performs automated tasks, but humans are involved at critical points to provide guidance, review and validate results, and make decisions based on their expertise and judgment. This helps ensure that the AI's outputs are reliable, ethical, and aligned with human expectations (more details in the next section).

On the opposite point of the spectrum of HitL is Human in Command, also known as MitL [12, 44, 7], in which the machine enters in the pipeline of human reasoning. There is a kind of augmented intelligence, also known as intelligence amplification, focusing on enhancing human intelligence with AI technologies. In this scenario, AI systems assist humans in complex tasks, providing data analysis, recommendations, or automation to support decision-making. The human remains in control and leverages the AI system as a tool for better efficiency and effectiveness. As in HitL, the presence of cooperation is also notorious here, but this time it flows from the machine to the human.

In the middle of the spectrum there is place for negotiation and equal involvement of machines and humans. In this case, the mixed-decision is taken by processes in which there might be a flat architecture, with no prevalence of control from one part. Thus, there is a kind of equilibrium in the policy power.

5.2 Active Learning

In the context of HitL described in the previous section, when the machine's task involves learning, active learning comes into play, as opposed to the conventional passive learning where the machine receives fully labeled datasets. Active learning [59] can indeed be viewed as a subset of HitL. However, it is important to note that, while most situations in active learning traditionally assume humans as the oracle, there is an openness to include other artificial agents also as oracles. Therefore, active learning is considered a form of HitL only when a human serves as the oracle, specifically in the context of annotation tasks. Hence, the segment of active learning involving a non-human oracle naturally falls beyond the purview of HitL. In general, active learning is considered a ML technique that aims to improve model performance and reduce labeling efforts by selectively and strategically choosing the most informative instances for which to request human annotation. It is proposed as part of the solution of a problem of traditional supervised learning, where a large

labeled dataset is used to train a model. In this setting, acquiring labeled data can be costly and time-consuming, especially when experts are needed for annotation. Active learning addresses this challenge by actively selecting data samples that are likely to be the most valuable or uncertain for model training. This serves as a reasonable justification for labeling active learning agents as curious agents.[59].

The selection of instances for annotation can be based on different strategies, such as uncertainty sampling (least confidence sampling, margin of confidence sampling, entropy-based sampling, ensemble-based sampling, etc.), or diversity sampling [59, 46]. These strategies aim to choose instances that are expected to have the greatest impact on improving the model's performance.

By actively selecting informative instances for annotation, active learning reduces the number of labeled examples required compared to traditional, often passive, supervised learning, while maintaining or even improving model accuracy. It enables efficient use of limited labeling resources by focusing on the most relevant instances.

Active learning finds applications in various domains, including text classification, image recognition, speech processing, and bioinformatics, where labeled data is expensive or time-consuming to obtain. It is a powerful tool for accelerating the annotation process and enhancing the performance of ML models.

5.3 Learning from Demonstration

As AI agents take on more complex real-world tasks, a major challenge is how to enable them to learn skills effectively. Traditional reinforcement learning methods [67] that rely solely on rewards from the environment can be inefficient, unsafe, or infeasible for acquiring complex behaviors. This has led to growing interest in leveraging human guidance to accelerate and direct ML [13]. In the light of human-AI cooperation, approaches that facilitate learning from observing human behaviour, including carefully selected human demonstrations, critiques, and preferences, is becoming increasingly important [14]. This section explores the related paradigms of learning from demonstration, learning from observation, imitation learning, apprenticeship learning, inverse reinforcement learning, and reinforcement learning from human feedback [63, 13, 25, 65, 1, 2, 6, 14, 29, 30, 39, 61].

There is no consensus in the taxonomy of these concepts, mainly at the topmost concept. Typically, the broader concept, learning from demonstration (also called imitation learning or apprenticeship learning [56]), is split into a direct method involving behaviour cloning (direct imitation learning) and an indirect method called inverse reinforcement learning (indirect imitation learning). Russell and Norvig [55] consider the concepts of imitation learning and inverse reinforcement learning under the umbrella of apprenticeship learning (corresponding to the terms of learning from demonstration of other approaches; programming by demonstration is another term also used).

Direct methods (imitation learning in the approach of Russell and Norvig) provide algorithms to allow mimicking expert behaviors, going beyond just memorization

to generalizing appropriately. Behavioral cloning is a common technique where supervised learning on state-action mappings reproduces demonstrated policies. However, directly copied policies tend to fail if the agent encounters new situations. More advanced methods aim to recover from errors and handle distributional shifts at test time.

Indirect methods are applied to surpass rote imitation. This is the case of inverse reinforcement learning which is a technique used to infer the hidden rewards and preferences driving the expert's choice of actions, i.e., the idea is to infer the hidden reward function (reward model) or intention behind demonstrations that explain expert behavior [75], which can then be used for policy improvement. This allows going beyond surface imitation to finding the underlying objectives. By uncovering the underlying objectives optimized by the teacher's demonstrations, the agent can achieve improved performance on its own through reinforcement learning. However, inverse reinforcement learning remains susceptible to ambiguities and depends heavily on high-quality demonstrations.

In general, learning from demonstration develops the core technical framework for agents to mimic observed behaviors. Advances in this area of AI translate to improved capabilities for leveraging human guidance. Combining it with reinforcement learning is a promising direction for future research. In traditional reinforcement learning, an agent learns by receiving rewards or penalties from the environment based on its actions. However, designing accurate reward functions for complex tasks can be difficult, and reinforcement learning agents may struggle to learn optimal behaviors. A recent advancement in reinforcement learning known as reinforcement learning from human feedback has garnered significant attention (e.g., in $chatGPT-3.5$ [51]). This technique combines the strengths of reinforcement learning and human expertise, allowing human feedback to guide and accelerate the learning process of AI systems. Thus, it might be considered a subfield of reinforcement learning in which an AI agent learns to perform tasks through interaction with human-generated feedback instead of a traditional reward signal provided by the environment. It aims to bridge the gap between traditional reinforcement learning algorithms and human guidance to make the learning process more efficient and effective, especially when defining reward functions is challenging or costly [29]. This collaborative approach not only enhances the performance of AI algorithms but also opens up new possibilities for human-AI collaboration in solving complex real-world challenges. This feedback can come in various forms such as:

- Comparison-based feedback: Humans provide preferences between different trajectories or actions, indicating which one is better. The agent uses this feedback to learn a policy that aligns with human preferences.
- Reward modeling: Humans provide feedback in the form of reward values for different states or actions. The agent learns a reward model from this feedback and then uses it to optimize its policy.
- Corrective feedback: Humans provide corrective feedback when the agent's actions are suboptimal or incorrect. The agent adjusts its policy based on this feedback.

Reinforcement learning from human feedback has applications in various domains, including robotics [34], gaming, natural language processing and autonomous systems. As aforementioned, it has also been used in the long process of training *chatGPT 3.5*. It helps address challenges such as sample efficiency, safety, and rapid adaptation to new tasks. While reinforcement learning from human feedback offers advantages, it also poses challenges related to human bias, inconsistency in feedback, and the scalability of human involvement. Researchers are actively exploring techniques to effectively integrate human feedback into reinforcement learning algorithms to improve learning performance and reduce the reliance on handcrafted reward functions. The primary goal of this ongoing research is to consistently combine the strengths of reinforcement learning and human guidance for increasingly powerful and reliable learning systems.

5.4 Trustworthy and Responsible Artificial Intelligence

Human-AI cooperation recognizes the importance of human expertise and judgment, and seeks to create synergistic partnerships between AI and human agents to achieve more effective, trustworthy, and beneficial outcomes. Trustworthiness is actually an important aspect in this binary relation scenario human-artificial agents, nowadays. This binary system may not work well without it.

The quest for trustworthy and responsible AI is a multifaceted endeavor encompassing technical, ethical, legal, and social dimensions. Achieving this goal necessitates collaborative efforts from academia, industry, policymakers, and society at large. As AI advances, the emphasis on developing AI systems that are not only powerful but also reliable and ethical remains crucial for unlocking AI's potential benefits while mitigating its risks.

While various ethical concerns related to AI underscore the need for addressing a wide range of topics, trust is often regarded as essential for AI development. According to the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) [31], a trusted AI comprises three essential components, legality, ethics, and robustness, which should ideally work harmoniously. This means that trusted AI must comply with all laws and regulations ethically, ensuring technological and social integrity. However, implementing these components may not always be straightforward or justifiable. For instance, when facial recognition is employed to identify criminals through images, ethical considerations like an individual's right to privacy may conflict.

The fundamental principles forming the foundation of these components are respect for human autonomy, prevention of harm, justice, and explainability. Despite varying interpretations of ethical and moral behavior, the European Commission's AI HLEG [31] defined these principles as reflections of inherent human rights. These rights encompass human dignity, freedom, equality, non-discrimination, solidarity, democracy, justice, and the rule of law.

Human autonomy is crucial for preserving dignity and freedom. Prunkl [54] outlines two aspects of autonomy, authenticity and agency, which emphasize humans' control over decisions and actions in the presence of AI. Respect for agency ensures that AI should empower humans to make decisions aligned with their beliefs and values, without external influences.

The prevention of harm extends beyond physical harm to include mental well-being and environmental impacts. It also emphasizes the importance of robust AI to thwart malicious use. Equity, another key principle, advocates for unbiased AI systems and procedural fairness in decision-making. This includes accountability and the right to challenge AI-generated decisions.

Explainability plays a pivotal role in building trust between humans and AI, especially for complex algorithms whose inner workings are opaque to non-experts. This transparency is crucial, particularly in high-stakes domains like healthcare.

While these principles provide a framework for trustworthy AI, the European Commission's AI HLEG outlines seven requirements for implementing these principles effectively:

1. Human agency and oversight: Before developing an AI system, potential violations of human rights must be thoroughly assessed. Throughout the development process, considerations for human autonomy should be integrated at each stage. Ultimately, there should be a degree of human involvement in decision-making processes, either as a mandatory requirement for each decision or as an optional capability.
2. Technical robustness and safety: AI systems must exhibit robustness primarily from a technical standpoint. Contingency plans should be in place to address security breaches. Additionally, AI systems must demonstrate accuracy, reliability, and reproducibility. In this context, reliable AI is defined as one that operates effectively across a variety of inputs and scenarios, as articulated by the AI HLEG.
3. Privacy and data management: All data inputted into AI systems, as well as any data generated, must be safeguarded and kept confidential, accessible only to authorized and qualified personnel. Data quality must be verified, and its integrity must be preserved throughout processing and storage.
4. Transparency: AI systems should be transparent, ensuring that all decisions and functionalities are well-documented to enhance accountability. AI systems should also be explainable, allowing users to understand the rationale behind AI-generated decisions. Furthermore, users should have the option to choose between interacting with AI systems or humans, necessitating clear labeling of AI agents to prevent deception.
5. Diversity, non-discrimination, and fairness: AI systems must be designed to avoid biases through meticulous evaluation of datasets and adherence to system development requirements. The usability of AI systems should be intuitive and accessible to all users. Continuous feedback from stakeholders should be sought throughout the AI's lifecycle to ensure inclusivity and fairness.
6. Social and environmental well-being: The impact of AI on the environment, society, and democratic principles requires close monitoring. Ethical considerations

should extend to environmental impacts, such as energy consumption and waste generation resulting from increased demand for electronic goods.
7. Accountability: Regular internal and external audits of AI systems must be conducted, with the results available for scrutiny. This is crucial, particularly in cases of negative impacts, where documentation and sharing of incidents aim to prevent similar occurrences in the future. Adequate compensation should be provided to those affected by such incidents to ensure fairness and accountability.

5.5 Summary

In this section, we focused on those multi-agent systems which include human agents, giving rise to forms of cooperation between humans and artificial agents. The concept of cooperative AI was presented as well as aspects of the relations between humans and artificial agents, naturally calling for the concepts of HitL, MitL (Human-in-Command), and the related spectrum of autonomy and control of the artificial and human agents. The roles of humans in those multi-agent systems, as key players in the design, development, deployment of AI systems, was presented. In fact, humans are the primary authors/creators of AI systems, although other AI agents may also be admitted to do it in collaboration with humans eventually and ultimately alone (e.g., Meta ML). Humans may define knowledge structures, the sensors, the algorithms of decision-making and learning, etc., of artificial agents. In addition, humans are also sources of data for training ML algorithms, and the beneficiaries of the outputs of the AI systems. All those phases of design, development, and deployment should be conceived with the guidelines, possibly under regulations, that guarantee that the AI system is designed for the beneficial of humans, and the agency of these should be respected. A key parameter is the level of autonomy of the AI system. A contribution from both humans and artificial agents should be considered, in some situations giving more autonomy to the artificial system, in others less. The question would always be to provide a reasonable balance so that the symbiosis happens.

6 Discussion: past, present, challenges, opportunities, and the future AI

The historical progression of intelligent agents and AI has been an intriguing journey. AI traces its roots to the mid-20th century, starting with rule-based systems and symbolic reasoning. However, early AI systems were constrained in their ability to deal with real-world complexity. These limitations led to periods of soaring expectations followed by "AI winters" characterized by slowed progress due to unmet hopes.

Today, we find ourselves in an era marked by swift advancements in AI, thanks to ML and, in particular, deep learning. This breakthrough has transformed the land-

scape of AI, enabling capabilities such as image recognition, language translation, and game-playing to match or surpass human performance levels. Intelligent agents, exemplified by virtual assistants like Siri and Alexa, self-driving cars, and recommendation systems, are now an integral part of our daily lives. AI algorithms are used to analyze vast amounts of data about our preferences, behaviors, and interactions, and provide us with personalized recommendations for products, services, and content. This has transformed the way we shop, consume media, and interact with the world around us. AI-powered robots, drones, and vehicles are increasingly being used in various industries, from manufacturing and logistics to healthcare and agriculture. These systems can perform tasks that are dangerous, repetitive, or require high precision, freeing up humans to focus on more creative and complex tasks. AI technologies such as speech recognition and natural language processing have made it possible for us to interact with machines using our voice and language, enabling new forms of communication and access to information. AI algorithms can also analyze large datasets to identify patterns, trends, and insights that humans may not be able to detect. This has applications in various fields, from finance and marketing to healthcare and social sciences. AI technologies are increasingly being designed to work alongside humans, augmenting their capabilities and expertise. This has the potential to improve decision-making, creativity, and problem-solving in various domains.

In spite of these advances, intelligent agents confront a range of challenges and limitations. One challenge pertains to their reliance on extensive labeled datasets, limiting their adaptability and necessitating continual data updates. Another limitation is their struggle with comprehending common sense and reasoning akin to humans, which affects their adaptability and contextual understanding. Ethical concerns have emerged, including biases in AI algorithms, privacy issues, and the potential for misuse, raising significant societal and ethical questions. Furthermore, AI still grapples with creativity and abstract thinking, excelling in tasks necessitating pattern recognition, but faltering when confronted with abstract concepts. Moreover, the complexity of many AI models renders them opaque, making it challenging to discern their decision-making processes.

Looking back, it is evident that a more substantial emphasis on interdisciplinary research and collaboration with cognitive science could have advanced AI development. Ethical considerations and long-term consequences should have been foundational from the outset.

The gap between artificial and human intelligence remains substantial, as AI lacks the comprehensive understanding, common sense, emotional intelligence, and creativity intrinsic to human intelligence.

Looking ahead, the future of intelligent agents is brimming with potential. It includes the pursuit of general AI, with AI systems possessing broader learning capabilities and common-sense reasoning. Collaborative endeavors between humans and AI offer the opportunity to augment human abilities in fields like healthcare, education, and research. Ethical AI development is indispensable, and AI safety is pivotal to ensure AI systems are robust against unintended consequences.

The importance of embracing intelligent agents cannot be overstated, but responsible adoption is equally critical. It is imperative to consider ethical principles, privacy concerns, and societal implications in AI utilization [20].

Generative AI has the potential to yield both positive and negative consequences in the future. Challenges may arise from deepfakes, misinformation, and the misuse of AI in cyberattacks, necessitating robust regulation, education, and technological safeguards.

AI is undergoing a renaissance sparked by explosive growth in data, advancements in deep learning algorithms, and expanded computational power through cloud computing and specialized hardware. The confluence of these factors is causing a Cambrian explosion of creative new approaches and capabilities.

In computer vision, techniques like convolutional and capsule networks have revolutionized image classification, object detection, and semantic segmentation. Generative adversarial networks can synthesize remarkably realistic photos, videos, and voices, a feat once considered squarely in the realm of science fiction.

In natural language, vast pretrained models such as GPT display impressive language generation abilities and even glimmers of common sense, while transformer-based models have achieved startling improvements on translation, question answering, and other natural language programming tasks. Chatbots like ChatGPT, Claude, Bing can carry out remarkably human-like conversations thanks to progress in dialogue systems.

Reinforcement learning has enabled superhuman gameplay in Go, chess, StarCraft II, and other arenas by extending deep learning to sequential decision making under uncertainty. Meanwhile, robotic capabilities in dexterity and mobility are expanding dramatically by incorporating deep learning, better 3D sensing, and novel hardware systems.

We are witnessing not just quantitative improvements in narrow applications but surprising qualitative leaps into emergent intelligence. Multi-modal, self-supervised learning promises to take this progress even further by developing more flexible, human-like learning algorithms. As barriers in computational power, algorithmic innovation, and data availability are overcome, the pace of advancement promises to be breathtaking. The future of AI holds tremendous excitement, promise, and perhaps some trepidation when envisioning where exponential trends may lead.

In conclusion, the journey of AI agents has been a testament to human ingenuity, with both triumphs and tribulations. Ethical considerations and responsible development should guide the AI landscape. Perhaps the solution lies within AI itself. If algorithms are powerful enough to build robust and reliable AI systems, why not leverage that power to imbue moral and ethical principles? Much of what prevents humans from harmful behaviors can be attributed to the ethical dimensions ingrained in our thinking. As such, imparting similar capacities for ethical reasoning in AI could significantly aid safety and trust. However, this remains an immense challenge. Human morality develops through complex sociocultural learning – an arduous process to replicate artificially. Significant innovations are required for AI to grapple with competing ethical frameworks and make contextually appropriate judgments of right and wrong. Success in this endeavor could greatly advance human

benefits, but progress will require continued research and responsible implementation. In the interim, humans have an obligation to remain vigilant about risks and limitations to ensure AI technologies are developed and used for moral good. However, this vigilance must not be conflated with prohibiting AI research. Perhaps even more efforts should be employed on developing new algorithms and applications, so that beneficial implementations of AI can outpace harmful ones. Retaining openness to AI's remarkable potential is essential. With ethical progress on all fronts, humans and AI agents can meaningfully cooperate in tackling the most pressing problems that confront society. Humans must be vigilant about potential risks and limitations AI while also fearlessly embracing the amazing possibilities it offers. We are decidedly entering a new era, one whose dawn rays are just peeking over the horizon.

References

1. Abbeel, P., Ng, A.Y.: Apprenticeship learning via inverse reinforcement learning. Proceedings of the twenty-first international conference on Machine learning (2004)
2. Argall, B.D., Chernova, S., Veloso, M., Browning, B.: A survey of robot learning from demonstration. Robotics and autonomous systems **57**(5), 469–483 (2009)
3. Bach, J.: Principles of synthetic intelligence. Oxford University Press (2008)
4. Bahrammirzaee, A.: A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Computing and Applications **19**(8), 1165–1195 (2010)
5. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S.: Quantum machine learning. Nature **549**(7671), 195–202 (2017)
6. Billard, A., Calinon, S., Dillmann, R., Schaal, S.: Robot programming by demonstration. Springer handbook of robotics pp. 1371–1394 (2008)
7. Bradshaw, J.M., Feltovich, P.J., Johnson, M.: Human-agent interaction. Handbook of Human-Machine Interaction pp. 283–302 (2011)
8. Bratman, M.E.: Intention, plans, and practical reason. Harvard University Press (1987)
9. Campbell, M., Hoane Jr, A.J., Hsu, F.H.: Deep blue. Artificial Intelligence **134**(1-2), 57–83 (2002)
10. Cassimatis, N.L., Trafton, J.G., Bugajska, M.D., Schultz, A.C.: Integrating cognitive models based on different computational methods. In: Proceedings of the National Conference on Artificial Intelligence. vol. 20, p. 1081 (2005)
11. Chanin, R., Espinosa, A., Demisse, B., Kley, J., Russell, J., BenYishay, A.: Applying artificial intelligence to combat environmental degradation and climate change: A landscape review of applications, tools, solutions, and challenges. Tech. rep., World Resources Institute (2022)
12. Chen, J., Barnes, M.J., Harper-Sciarini, M.: Human–agent collaboration for disaster relief. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **55**(1), 449–453 (2011)
13. Chernova, S., Thomaz, A.L.: Robot learning from human teachers. In: Synthesis Lectures on Artificial Intelligence and Machine Learning (2014), https://api.semanticscholar.org/CorpusID:26200231
14. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems **30** (2017)
15. Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K.R., Leibo, J.Z., Larson, K., Graepel, T.: Open problems in cooperative ai. arXiv preprint arXiv:2012.08630 (2020)

16. Dahl, E.A.: From logic to learning: Representing abstract knowledge in neural networks and learning symbolic inference. In: International Conference on Engineering Psychology and Cognitive Ergonomics. pp. 247–256. Springer (2010)
17. Daugherty, P.R., Wilson, H.J., Chowdhury, R.: Myths and realities of ai. IEEE Intelligent systems **34**(4), 66–71 (2019)
18. Descartes, R.: Meditations on First Philosophy. Renaissance Classics (1641)
19. Deutsch, D.: Quantum theory, the church-turing principle and the universal quantum computer. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences **400**(1818), 97–117 (1985)
20. Dignum, V.: Responsible artificial intelligence. Springer International Publishing (2019)
21. Domingos, P.: The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books (2015)
22. Dunjko, V., Briegel, H.J.: Machine learning artificial intelligence in the quantum domain. Reports on Progress in Physics **81**(7), 074001 (2018)
23. Ferber, J.: Multi-agent systems: an introduction to distributed artificial intelligence. Addison-Wesley Longman Publishing Co., Inc. (1999)
24. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., et al.: Building watson: An overview of the deepqa project. In: AI magazine. vol. 31, pp. 59–79 (2010)
25. Finn, C., Levine, S., Abbeel, P.: Guided cost learning: Deep inverse optimal control via policy optimization. In: International Conference on Machine Learning. pp. 49–58. PMLR (2016)
26. Forbus, K.D., Usher, J.: The companions cognitive architecture. AI Magazine **23**(2), 41–41 (2002)
27. Gardner, H.: Frames of mind: The theory of multiple intelligences. Hachette UK (2011)
28. Gottfredson, L.S.: Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. Intelligence **24**(1), 13–23 (1997)
29. Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S.J., Dragan, A.: Inverse reward design. In: Advances in neural information processing systems. pp. 6768–6777 (2017)
30. Herman, M., Gindele, T., Wagner, J., Schmitt, F., Burgard, W.: The tracking-learning-detection method for direct training of resource-constrained systems. IEEE Robotics and Automation Letters **2**(4), 2566–2573 (2017)
31. High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI (2019). https://doi.org/https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
32. Holzinger, A.: Explainable ai and multi-modal causability in medicine. I Com **19**, 171–179 (2021). https://doi.org/10.1515/icom-2020-0024
33. Huhns, M.N., Singh, M.P.: Distributed artificial intelligence, vol. 2. Pitman London (1998)
34. Knox, W.B., Stone, P.: Reinforcement learning from human reward: Discounting in episodic tasks. In: 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication. pp. 878–885. IEEE (2012)
35. Kurzweil, R.: The age of intelligent machines. MIT press (1990)
36. Laird, J.E., Newell, A., Rosenbloom, P.S.: Soar: An architecture for general intelligence. Artificial intelligence **33**(1), 1–64 (1987)
37. Legg, S., Hutter, M.: Universal intelligence: A definition of machine intelligence. Minds and machines **17**(4), 391–444 (2007)
38. Leibniz, G.W.: Monadology. Oxford University Press (1714)
39. Loftin, R., Lajoie, M., Hill, D., Phillips, A.: A strategy for using crowdsourcing to improve machine learning classifiers for sentiment analysis. Proceedings of the Annual Meeting of the Cognitive Science Society **36**(36) (2014)
40. Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A.: Explainable artificial intelligence: Concepts, applications, research challenges and visions. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 1–16. Springer International Publishing, Cham (2020)

41. Manyika, J., Chui, M., Miremadi, M., Bughin, J., George, K., Willmott, P., Dewhurst, M.: The social and economic implications of artificial intelligence technologies. McKinsey Global Institute (2017)
42. Mayor, A.: Gods and Robots: Myths, Machines, and Ancient Dreams of Technology. Princeton University Press (2018). https://doi.org/10.2307/j.ctvc779xn
43. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics **5**(4), 115–133 (1943)
44. Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D., Procci, K.: Intelligent agent transparency in human–agent teaming for multi-uxv management. Human factors **58**(3), 401–415 (2016)
45. Mitchell, T.M.: The discipline of machine learning (2006)
46. Munro, R.M.: Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI. Manning Publications Co. (2021)
47. Neisser, U., Boodoo, G., Bouchard Jr, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., et al.: Intelligence: knowns and unknowns. American psychologist **51**(2), 77–101 (1996)
48. von Neumann, J., Wiener, N.: A certain statistical method applicable to groups of organisms. Annals of Mathematical Statistics **19**(4), 357–366 (1948)
49. Nilsson, N.J.: Human-level artificial intelligence? be serious! AI magazine **26**(4), 68–75 (2005)
50. Nilsson, N.J.: The Quest for Artificial Intelligence. Cambridge University Press (2009)
51. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Miller, Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.J.: Training language models to follow instructions with human feedback. ArXiv **abs/2203.02155** (2022), https://api.semanticscholar.org/CorpusID:246426909
52. Parunak, H.V.D.: Multi-agent machine learning: A reinforcement approach. John Wiley Sons, Inc. (1997)
53. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (1988)
54. Prunkl, C.: Human autonomy in the age of artificial intelligence. Nature Machine Intelligence **4**, 99–101 (2022). https://doi.org/10.1038/s42256-022-00449-9
55. Russell, S.: Human compatible: Artificial intelligence and the problem of control. Penguin (2019)
56. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson (2020)
57. Santos, F.M.F.d.: Empowering Classical AI with Quantum Computing. Master's thesis (2022), http://hdl.handle.net/10316/102180
58. Searle, J.R.: Minds, brains, and programs. Behavioral and brain sciences **3**(3), 417–424 (1980)
59. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
60. Shannon, C.E.: Programming a computer for playing chess. Philosophical Magazine **41**(314), 256–275 (1950)
61. Shivaswamy, P., Bhattacharyya, C.: Coactive learning. In: Artificial Intelligence and Statistics. pp. 147–155. PMLR (2015)
62. Shoham, Y., Leyton-Brown, K.: Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press (2008)
63. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. nature **529**(7587), 484–489 (2016)
64. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering the game of go without human knowledge. Nature **550**(7676), 354–359 (2017)
65. Stadie, B.C., Levine, S., Abbeel, P.: Third-person imitation learning. In: International Conference on Learning Representations (2017)
66. Sternberg, R.J. (ed.): Handbook of intelligence. Cambridge University Press (2000)
67. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)

68. Swade, D.: The Difference Engine: Charles Babbage and the Quest to Build the First Computer. Viking Adult (2001)
69. Turing, A.M.: On computable numbers, with an application to the entscheidungsproblem. Proceedings of the London Mathematical Society **42**(1), 230–265 (1936)
70. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236), 433–460 (1950)
71. Weiss, G.: Multiagent systems. MIT press (2013)
72. Wittek, P.: Quantum machine learning: what quantum computing means to data mining. Academic Press (2014)
73. Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley & Sons (2009)
74. Yu, K.H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare. Nature Biomedical Engineering **2**(10), 719–731 (2018)
75. Ziebart, B.D., Maas, A., Bagnell, J.A., Dey, A.K.: Maximum entropy inverse reinforcement learning. In: Aaai. vol. 8, pp. 1433–1438 (2008)