Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans. Leading AI textbooks define the field as the study of "intelligent agents": any system that perceives its environment and takes actions that maximize its chance of achieving its goals.[a] Some popular accounts use the term "artificial intelligence" to describe machines that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving", however, this definition is rejected by major AI researchers.[b]

AI applications include advanced web search engines (e.g., Google), recommendation systems (used by YouTube, Amazon and Netflix), understanding human speech (such as Siri and Alexa), self-driving cars (e.g., Tesla), automated decision-making and competing at the highest level in strategic game systems (such as chess and Go).[2][citation needed] As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect.[3] For instance, optical character recognition is frequently excluded from things considered to be AI,[4] having become a routine technology.[5]

Artificial intelligence was founded as an academic discipline in 1956, and in the years since has experienced several waves of optimism,[6][7] followed by disappointment and the loss of funding (known as an "AI winter"),[8][9] followed by new approaches, success and renewed funding.[7][10] AI research has tried and discarded many different approaches since its founding, including simulating the brain, modeling human problem solving, formal logic, large databases of knowledge and imitating animal behavior. In the first decades of the 21st century, highly mathematical statistical machine learning has dominated the field, and this technique has proved highly successful, helping to solve many challenging problems throughout industry and academia.[11][10]

The various sub-fields of AI research are centered around particular goals and the use of particular tools. The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception, and the ability to move and manipulate objects.[c] General intelligence (the ability to solve an arbitrary problem) is among the field's long-term goals.[12] To solve these problems, AI researchers have adapted and integrated a wide range of problem-solving techniques -- including search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, probability and economics. AI also draws upon computer science, psychology, linguistics, philosophy, and many other fields.

The field was founded on the assumption that human intelligence "can be so precisely described that a machine can be made to simulate it".[d] This raises philosophical arguments about the mind and the ethics of creating artificial beings endowed with human-like intelligence. These issues have been explored by myth, fiction, and philosophy since antiquity.[14] Science fiction and futurology have also suggested that, with its enormous potential and power, AI may become an existential risk to humanity.[15][16]

Contents

1 History

2 Goals

2.1 Reasoning, problem solving

2.2 Knowledge representation

2.3 Planning

2.4 Learning

2.5 Natural language processing

2.6 Perception

2.7 Motion and manipulation

2.8 Social intelligence

2.9 General intelligence

3 Tools

3.1 Search and optimization

3.2 Logic

3.3 Probabilistic methods for uncertain reasoning

3.4 Classifiers and statistical learning methods

3.5 Artificial neural networks

3.6 Specialized languages and hardware

4 Applications

5 Philosophy

5.1 Defining artificial intelligence

5.2 Evaluating approaches to AI

5.3 Machine consciousness, sentience and mind

6 Future

6.1 Superintelligence

6.2 Risks

6.3 Ethical machines

6.4 Regulation

7 In fiction

8 See also

9 Explanatory notes

10 Citations

11 References

11.1 AI textbooks

11.2 History of AI

11.3 Other sources

12 Further reading

13 External links

14 Sources

History

Main articles: History of artificial intelligence and Timeline of artificial intelligence

Silver didrachma from Crete depicting Talos, an ancient mythical automaton with artificial intelligence

Artificial beings with intelligence appeared as storytelling devices in antiquity,[17] and have been common in fiction, as in Mary Shelley's Frankenstein or Karel Čapek's R.U.R.[18] These characters and their fates raised many of the same issues now discussed in the ethics of artificial intelligence.[19]

The study of mechanical or "formal" reasoning began with philosophers and mathematicians in antiquity. The study of mathematical logic led directly to Alan Turing's theory of computation, which suggested that a machine, by shuffling symbols as simple as "0" and "1", could simulate any conceivable act of mathematical deduction. This insight that digital computers can simulate any process of formal reasoning is known as the Church–Turing thesis.[20]

The Church-Turing thesis, along with concurrent discoveries in neurobiology, information theory and cybernetics, led researchers to consider the possibility of building an electronic brain.[21] The first work that is now generally recognized as AI was McCullouch and Pitts' 1943 formal design for Turing-complete "artificial neurons".[22]

When access to digital computers became possible in the mid-1950s, AI research began to explore the possibility that human intelligence could be reduced to step-by-step symbol manipulation, known as Symbolic AI or GOFAI. Approaches based on cybernetics or artificial neural networks were abandoned or pushed into the background.

The field of AI research was born at a workshop at Dartmouth College in 1956.[e][25] The attendees became the founders and leaders of AI research.[f] They and their students produced programs that the press described as "astonishing":[g] computers were learning checkers strategies, solving word problems in algebra, proving logical theorems and speaking English.[h][27] By the middle of the 1960s, research in the U.S. was heavily funded by the Department of Defense[28] and laboratories had been established around the world.[29]

Researchers in the 1960s and the 1970s were convinced that symbolic approaches would eventually succeed in creating a machine with artificial general intelligence and considered this the goal of their field.[30] Herbert Simon predicted, "machines will be capable, within twenty years, of doing any work a man can do".[31] Marvin Minsky agreed, writing, "within a generation ... the problem of creating 'artificial intelligence' will substantially be solved".[32]

They failed to recognize the difficulty of some of the remaining tasks. Progress slowed and in 1974, in response to the criticism of Sir James Lighthill[33] and ongoing pressure from the US Congress to fund more productive projects, both the U.S. and British governments cut off exploratory research in AI. The next few years would later be called an "AI winter", a period when obtaining funding for AI projects was difficult. [8]

In the early 1980s, AI research was revived by the commercial success of expert systems,[34] a form of AI program that simulated the knowledge and analytical skills of human experts. By 1985, the market for AI had reached over a billion dollars. At the same time, Japan's fifth generation computer project inspired the U.S and British governments to restore funding for academic research.[7] However, beginning with the collapse of the Lisp Machine market in 1987, AI once again fell into disrepute, and a second, longer-lasting winter began.[9]

Many researchers began to doubt that the symbolic approach would be able to imitate all the processes of human cognition, especially perception, robotics, learning and pattern recognition. A number of researchers began to look into "sub-symbolic" approaches to specific AI problems.[35] Robotics researchers, such as Rodney Brooks, rejected symbolic AI and focused on the basic engineering problems that would allow robots to move, survive, and learn their environment.[i] Interest in neural networks and "connectionism" was revived by Geoffrey Hinton, David Rumelhart and others in the middle of the 1980s.[40] Soft computing tools were developed in the 80s, such as neural networks, fuzzy systems, Grey system theory, evolutionary computation and many tools drawn from statistics or mathematical optimization.

AI gradually restored its reputation in the late 1990s and early 21st century by finding specific solutions to specific problems. The narrow focus allowed researchers to produce verifiable results, exploit more mathematical methods, and collaborate with other fields (such as statistics, economics and mathematics).[41] By 2000, solutions developed by AI researchers were being widely used, although in the 1990s they were rarely described as "artificial intelligence".[11]

Faster computers, algorithmic improvements, and access to large amounts of data enabled advances in machine learning and perception; data-hungry deep learning methods started to dominate accuracy benchmarks around 2012.[42] According to Bloomberg's Jack Clark, 2015 was a landmark year for artificial intelligence, with the number of software projects that use AI within Google increased from a "sporadic usage" in 2012 to more than 2,700 projects.[j] He attributes this to an increase in affordable neural networks, due to a rise in cloud computing infrastructure and to an increase in research tools and datasets.[10] In a 2017 survey, one in five companies reported they had "incorporated AI in some offerings or processes".[43] The amount of research into AI (measured by total publications) increased by 50% in the years 2015–2019.[44]

Numerous academic researchers became concerned that AI was no longer pursuing the original goal of creating versatile, fully intelligent machines. Much of current research involves statistical AI, which is overwhelmingly used to solve specific problems, even highly successful techniques such as deep learning. This concern has led to the subfield artificial general intelligence (or "AGI"), which had several well-funded institutions by the 2010s.[12]

Goals

The general problem of simulating (or creating) intelligence has been broken down into sub-problems. These consist of particular traits or capabilities that researchers expect an intelligent system to display. The traits described below have received the most attention.[c]

Reasoning, problem solving

Early researchers developed algorithms that imitated step-by-step reasoning that humans use when they solve puzzles or make logical deductions.[45] By the late 1980s and 1990s, AI research had developed methods for dealing with uncertain or incomplete information, employing concepts from probability and economics.[46]

Many of these algorithms proved to be insufficient for solving large reasoning problems because they experienced a "combinatorial explosion": they became exponentially slower as the problems grew larger.[47] Even humans rarely use the step-by-step deduction that early AI research could model. They solve most of their problems using fast, intuitive judgments.[48]

An ontology represents knowledge as a set of concepts within a domain and the relationships between those concepts.

Knowledge representation

Main articles: Knowledge representation, Commonsense knowledge, Description logic, and Ontology

Knowledge representation and knowledge engineering[49] allow AI programs to answer questions intelligently and make deductions about real world facts.

A representation of "what exists" is an ontology: the set of objects, relations, concepts, and properties formally described so that software agents can interpret them.[50] The most general ontologies are called upper ontologies, which attempt to provide a foundation for all other knowledge and act as mediators between domain ontologies that cover specific knowledge about a particular knowledge domain (field of interest or area of concern). A truly intelligent program would also need access to commonsense knowledge; the set of facts that an average person knows. The semantics of an ontology is typically represented in a description logic, such as the Web Ontology Language.[51]

AI research has developed tools to represent specific domains, such as: objects, properties, categories and relations between objects;[51] situations, events, states and time;[52] causes and effects;[53] knowledge about knowledge (what we know about what other people know);.[54] default reasoning (things that humans assume are true until they are told differently and will remain true even when other facts are changing); [55] as well as other domains. Among the most difficult problems in AI are: the breadth of commonsense knowledge (the number of atomic facts that the average person knows is enormous);[56] and the sub-symbolic form of most commonsense knowledge (much of what people know is not represented as "facts" or "statements" that they could express verbally).[48]

Formal knowledge representations are used in content-based indexing and retrieval,[57] scene interpretation,[58] clinical decision support,[59] knowledge discovery (mining "interesting" and actionable inferences from large databases),[60] and other areas.[61]

Planning

Main article: Automated planning and scheduling

An intelligent agent that can plan makes a representation of the state of the world, makes predictions about how their actions will change it and makes choices that maximize the utility (or "value") of the available choices.[62] In classical planning problems, the agent can assume that it is the only system acting in the world, allowing the agent to be certain of the consequences of its actions.[63] However, if the agent is not the only actor, then it requires that the agent reason under uncertainty, and continuously re-assess its environment and adapt.[64] Multi-agent planning uses the cooperation and competition of many agents to achieve a given goal. Emergent behavior such as this is used by evolutionary algorithms and swarm intelligence.[65]

Learning

Main article: Machine learning

Machine learning (ML), a fundamental concept of AI research since the field's inception,[k] is the study of computer algorithms that improve automatically through experience.[l]

Unsupervised learning finds patterns in a stream of input. Supervised learning requires a human to label the input data first, and comes in two main varieties: classification and numerical regression. Classification is used to determine what category something belongs in -- the program sees a number of examples of things from several categories and will learn to classify new inputs. Regression is the attempt to produce a function that describes the relationship between inputs and outputs and predicts how the outputs should change as the inputs change. Both classifiers and regression learners can be viewed as "function approximators" trying to learn an unknown (possibly implicit) function; for example, a spam classifier can be viewed as learning a function that maps from the text of an email to one of two categories, "spam" or "not spam".[69] In reinforcement learning the agent is rewarded for good responses and punished for bad ones. The agent classifies its responses to form a strategy for operating in its problem space.[70] Transfer learning is when knowledge gained from one problem is applied to a new problem.[71]

Computational learning theory can assess learners by computational complexity, by sample complexity (how much data is required), or by other notions of optimization.[72]

A parse tree represents the syntactic structure of a sentence according to some formal grammar.

Natural language processing

Main article: Natural language processing

Natural language processing (NLP)[73] allows machines to read and understand human language. A sufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of NLP include information retrieval, question answering and machine translation.[74]

Symbolic AI used formal syntax to translate the deep structure of sentences into logic. This failed to produce useful applications, due to the intractability of logic[47] and the breadth of commonsense knowledge.[56] Modern statistical techniques include co-occurrence frequencies (how often one word appears near another), "Keyword spotting" (searching for a particular word to retrieve information), transformer-based deep learning (which finds patterns in text), and others.[75] They have achieved acceptable accuracy at the page or paragraph level, and, by 2019, could generate coherent text.[76]

Feature detection (pictured: edge detection) helps AI compose informative abstract structures out of raw data.

Perception

Main articles: Machine perception, Computer vision, and Speech recognition

Machine perception[77] is the ability to use input from sensors (such as cameras, microphones, wireless signals, and active lidar, sonar, radar, and tactile sensors) to deduce aspects of the world. Applications include speech recognition,[78] facial recognition, and object recognition.[79]

Computer vision is the ability to analyze visual input.[80]

Motion and manipulation

Main article: Robotics

AI is heavily used in robotics.[81] Localization is how a robot knows its location and map its environment. When given a small, static, and visible environment, this is easy; however, dynamic environments, such as (in endoscopy) the interior of a patient's breathing body, pose a greater challenge.[82]

Motion planning is the process of breaking down a movement task into "primitives" such as individual joint movements. Such movement often involves compliant motion, a process where movement requires maintaining physical contact with an object. Robots can learn from experience how to move efficiently despite the presence of friction and gear slippage.[83]

Kismet, a robot with rudimentary social skills[84]

Social intelligence

Main article: Affective computing

Affective computing is an interdisciplinary umbrella that comprises systems which recognize, interpret, process, or simulate human feeling, emotion and mood.[85] For example, some virtual assistants are programmed to speak conversationally or even to banter humorously; it makes them appear more sensitive to the emotional dynamics of human interaction, or to otherwise facilitate human–computer interaction. However, this tends to give naïve users an unrealistic conception of how intelligent existing computer agents actually are.[86]

Moderate successes related to affective computing include textual sentiment analysis and, more recently, multimodal sentiment analysis), wherein AI classifies the affects displayed by a videotaped subject.[87]

General intelligence

Main article: Artificial general intelligence

A machine with general intelligence can solve a wide variety of problems with a breadth and versatility similar to human intelligence. There are several competing ideas about how to develop artificial general intelligence. Hans Moravec and Marvin Minsky argue that work in different individual domains can be incorporated into an advanced multi-agent system or cognitive architecture with general intelligence.[88] Pedro Domingos hopes that there is a conceptually straightforward, but mathematically difficult, "master algorithm" that could lead to AGI.[89] Others believe that anthropomorphic features like an artificial brain[90] or simulated child development[m] will someday reach a critical point where general intelligence emerges.

Tools

Search and optimization

Main articles: Search algorithm, Mathematical optimization, and Evolutionary computation

Many problems in AI can be solved theoretically by intelligently searching through many possible solutions:[91] Reasoning can be reduced to performing a search. For example, logical proof can be viewed as searching for a path that leads from premises to conclusions, where each step is the application of an inference rule.[92] Planning algorithms search through trees of goals and subgoals, attempting to find a path to a target goal, a process called means-ends analysis.[93] Robotics algorithms for moving limbs and grasping objects use local searches in configuration space.[94]

Simple exhaustive searches[95] are rarely sufficient for most real-world problems: the search space (the number of places to search) quickly grows to astronomical numbers. The result is a search that is too slow or never completes. The solution, for many problems, is to use "heuristics" or "rules of thumb" that prioritize choices in favor of those more likely to reach a goal and to do so in a shorter number of steps. In some search methodologies heuristics can also serve to eliminate some choices unlikely to lead to a goal (called "pruning the search tree"). Heuristics supply the program with a "best guess" for the path on which the solution lies.[96] Heuristics limit the search for solutions into a smaller sample size.[97]

A particle swarm seeking the global minimum

A very different kind of search came to prominence in the 1990s, based on the mathematical theory of optimization. For many problems, it is possible to begin the search with some form of a guess and then refine the guess incrementally until no more refinements can be made. These algorithms can be visualized as blind hill climbing: we begin the search at a random point on the landscape, and then, by jumps or steps, we keep moving our guess uphill, until we reach the top. Other optimization algorithms are simulated annealing, beam search and random optimization.[98] Evolutionary computation uses a form of optimization search. For example, they may begin with a population of organisms (the guesses) and then allow them to mutate and recombine, selecting only the fittest to survive each generation (refining the guesses). Classic evolutionary algorithms include genetic algorithms, gene expression programming, and genetic programming.[99] Alternatively, distributed search processes can coordinate via swarm intelligence algorithms. Two popular swarm algorithms used in search are particle swarm optimization (inspired by bird flocking) and ant colony optimization (inspired by ant trails).[100]

Logic

Main articles: Logic programming and Automated reasoning

Logic[101] is used for knowledge representation and problem solving, but it can be applied to other problems as well. For example, the satplan algorithm uses logic for planning[102] and inductive logic programming is a method for learning.[103]

Several different forms of logic are used in AI research. Propositional logic[104] involves truth functions such as "or" and "not". First-order logic[105] adds quantifiers and predicates, and can express facts about objects, their properties, and their relations with each other. Fuzzy logic assigns a "degree of truth" (between 0 and 1) to vague statements such as "Alice is old" (or rich, or tall, or hungry), that are too linguistically imprecise to be completely true or false.[106] Default logics, non-monotonic logics and circumscription are forms of logic designed to help with default reasoning and the qualification problem.[55] Several extensions of logic have been designed to handle specific domains of knowledge, such as: description logics;[51] situation calculus, event calculus and fluent calculus (for representing events and time);[52] causal calculus;[53] belief calculus (belief revision); and modal logics.[54] Logics to model contradictory or inconsistent statements arising in multi-agent systems have also been designed, such as paraconsistent logics.[citation needed]

Probabilistic methods for uncertain reasoning

Main articles: Bayesian network, Hidden Markov model, Kalman filter, Particle filter, Decision theory, and Utility theory

Expectation-maximization clustering of Old Faithful eruption data starts from a random guess but then successfully converges on an accurate clustering of the two physically distinct modes of eruption.

Many problems in AI (in reasoning, planning, learning, perception, and robotics) require the agent to operate with incomplete or uncertain information. AI researchers have devised a number of powerful tools to solve these problems using methods from probability theory and economics.[107] Bayesian networks[108] are a very general tool that can be used for various problems: reasoning (using the Bayesian inference algorithm),[n][110] learning (using the expectation-maximization algorithm),[o][112] planning (using decision networks)[113] and perception (using dynamic Bayesian networks).[114] Probabilistic algorithms can also be used for filtering, prediction, smoothing and finding explanations for streams of data, helping perception systems to analyze processes that occur over time (e.g., hidden Markov models or Kalman filters).[114]

A key concept from the science of economics is "utility": a measure of how valuable something is to an intelligent agent. Precise mathematical tools have been developed that analyze how an agent can make choices and plan, using decision theory, decision analysis,[115] and information value theory.[116] These tools include models such as Markov decision processes,[117] dynamic decision networks,[114] game theory and mechanism design.[118]

Classifiers and statistical learning methods

Main articles: Classifier (mathematics), Statistical classification, and Machine learning

The simplest AI applications can be divided into two types: classifiers ("if shiny then diamond") and controllers ("if diamond then pick up"). Controllers do, however, also classify conditions before inferring actions, and therefore classification forms a central part of many AI systems. Classifiers are functions that use pattern matching to determine a closest match. They can be tuned according to examples, making them very attractive for use in AI. These examples are known as observations or patterns. In supervised learning, each pattern belongs to a certain predefined class. A class is a decision that has to be made. All the observations combined with their class labels are known as a data set. When a new observation is received, that observation is classified based on previous experience.[119]

A classifier can be trained in various ways; there are many statistical and machine learning approaches. The decision tree is the simplest and most widely used symbolic machine learning algorithm.[120] K-nearest neighbor algorithm was the most widely used analogical AI until the mid-1990s.[121] Kernel methods such as the support vector machine (SVM) displaced k-nearest neighbor in the 1990s.[122] The naive Bayes classifier is reportedly the "most widely used learner"[123] at Google, due in part to its scalability.[124] Neural networks are also used for classification.[125]

Classifier performance depends greatly on the characteristics of the data to be classified, such as the dataset size, distribution of samples across classes, the dimensionality, and the level of noise. Model-based classifiers perform well if the assumed model is an extremely good fit for the actual data. Otherwise, if no matching model is available, and if accuracy (rather than speed or scalability) is the sole concern, conventional wisdom is that discriminative classifiers (especially SVM) tend to be more accurate than model-based classifiers such as "naive Bayes" on most practical data sets.[126]

Artificial neural networks

Main articles: Artificial neural network and Connectionism

A neural network is an interconnected group of nodes, akin to the vast network of neurons in the human brain.

Neural networks[125] were inspired by the architecture of neurons in the human brain. A simple "neuron" N accepts input from other neurons, each of which, when activated (or "fired"), casts a weighted "vote" for or against whether neuron N should itself activate. Learning requires an algorithm to adjust these weights based on the training data; one simple algorithm (dubbed "fire together, wire together") is to increase the weight between two connected neurons when the activation of one triggers the successful activation of another. Neurons have a continuous spectrum of activation; in addition, neurons can process inputs in a nonlinear way rather than weighing straightforward votes.

Modern neural networks model complex relationships between inputs and outputs or and find patterns in data. They can learn continuous functions and even digital logical operations. Neural networks can be viewed a type of mathematical optimization — they perform a gradient descent on a multi-dimensional topology that was created by training the network. The most common training technique is the backpropagation algorithm.[127] Other learning techniques for neural networks are Hebbian learning ("fire together, wire together"), GMDH or competitive learning.[128]

The main categories of networks are acyclic or feedforward neural networks (where the signal passes in only one direction) and recurrent neural networks (which allow feedback and short-term memories of previous input events). Among the most popular feedforward networks are perceptrons, multi-layer perceptrons and radial basis networks.[129]

Deep learning

Representing Images on Multiple Layers of Abstraction in Deep Learning

Representing Images on Multiple Layers of Abstraction in Deep Learning[130]

Deep learning[131] uses several layers of neurons between the network's inputs and outputs. The multiple layers can progressively extract higher-level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces.[132] Deep learning has drastically improved the performance of programs in many important subfields of artificial intelligence, including computer vision, speech recognition, image classification[133] and others.

Deep learning often uses convolutional neural networks for many or all of its layers. In a convolutional layer, each neuron receives input from only a restricted area of the previous layer called the neuron's receptive field. This can substantially reduce the number of weighted connections between neurons,[134] and creates a hierarchy similar to the organization of the animal visual cortex.[135]

In a recurrent neural network the signal will propagate through a layer more than once;[136] thus, an RNN is an example of deep learning.[137] RNNs can be trained by gradient descent,[138] however long-term gradients which are back-propagated can "vanish" (that is, they can tend to zero) or "explode" (that is, they can tend to infinity), known as the vanishing gradient problem.[139] The long short term memory (LSTM) technique can prevent this in most cases.[140]

Specialized languages and hardware

Main articles: Programming languages for artificial intelligence and Hardware for artificial intelligence

Specialized languages for artificial intelligence have been developed, such as Lisp, Prolog, TensorFlow and many others. Hardware developed for AI includes AI accelerators and neuromorphic computing.

Applications

Main article: Applications of artificial intelligence

For this project the AI had to learn the typical patterns in the colors and brushstrokes of Renaissance painter Raphael. The portrait shows the face of the actress Ornella Muti, "painted" by AI in the style of Raphael.

AI is relevant to any intellectual task.[141] Modern artificial intelligence techniques are pervasive and are too numerous to list here.[142] Frequently, when a technique reaches mainstream use, it is no longer considered artificial intelligence; this phenomenon is described as the AI effect.[143]

In the 2010s, AI applications were at the heart of the most commercially successful areas of computing, and have become a ubiquitous feature of daily life. AI is used in search engines (such as Google Search), targeting online advertisements,[144] recommendation systems (offered by Netflix, YouTube or Amazon), driving internet traffic,[145][146] targeted advertising (AdSense, Facebook), virtual assistants (such as Siri or Alexa),[147] autonomous vehicles (including drones and self-driving cars), automatic language translation (Microsoft Translator, Google Translate), facial recognition (Apple's Face ID or Microsoft's DeepFace), image labeling (used by Facebook, Apple's iPhoto and TikTok) and spam filtering.

There are also thousands of successful AI applications used to solve problems for specific industries or institutions. A few examples are: energy storage,[148] deepfakes,[149] medical diagnosis, military logistics, or supply chain management.

Game playing has been a test of AI's strength since the 1950s. Deep Blue became the first computer chess-playing system to beat a reigning world chess champion, Garry Kasparov, on 11 May 1997.[150] In 2011, in a Jeopardy! quiz show exhibition match, IBM's question answering system, Watson, defeated the two greatest Jeopardy! champions, Brad Rutter and Ken Jennings, by a significant margin.[151] In March 2016, AlphaGo won 4 out of 5 games of Go in a match with Go champion Lee Sedol, becoming the first computer Go-playing system to beat a professional Go player without handicaps.[152] Other programs handle imperfect-information games; such as for poker at a superhuman level, Pluribus[p] and Cepheus.[154] DeepMind in the 2010s developed a "generalized artificial intelligence" that could learn many diverse Atari games on its own.[155]

By 2020, Natural Language Processing systems such as the enormous GPT-3 (then by far the largest artificial neural network) were matching human performance on pre-existing benchmarks, albeit without the system attaining commonsense understanding of the contents of the benchmarks.[156] DeepMind's AlphaFold 2 (2020) demonstrated the ability to determine, in hours rather than months, the 3D structure of a protein.[157] Other applications predict the result of judicial decisions,[158] create art (such as poetry or painting) and prove mathematical theorems.

Philosophy

Main article: Philosophy of artificial intelligence

Defining artificial intelligence

Thinking vs. acting: the Turing test

Main articles: Turing test, Dartmouth Workshop, and Synthetic intelligence

Alan Turing wrote in 1950 "I propose to consider the question 'can machines think'?"[159] He advised changing the question from whether a machine "thinks", to "whether or not it is possible for machinery to show intelligent behaviour".[160] The only thing visible is the behavior of the machine, so it does not matter if the machine is conscious, or has a mind, or whether the intelligence is merely a "simulation" and not "the real thing". He noted that we also don't know these things about other people, but that we extend a "polite convention" that they are actually "thinking". This idea forms the basis of the Turing test.[161][q]

Acting humanly vs. acting intelligently: intelligent agents

Main article: Intelligent agents

AI founder John McCarthy said: "Artificial intelligence is not, by definition, simulation of human intelligence".[163] Russell and Norvig agree and criticize the Turing test. They wrote: "Aeronautical engineering texts do not define the goal of their field as 'making machines that fly so exactly like pigeons that they can fool other pigeons.'"[162] Other researchers and analysts disagree and have argued that AI should simulate natural intelligence by studying psychology or neurobiology.[r]

The intelligent agent paradigm[165] defines intelligent behavior in general, without reference to human beings. An intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success. Any system that has goal-directed behavior can be analyzed as an intelligent agent: something as simple as a thermostat, as complex as a human being, as well as large systems such as firms, biomes or nations. The intelligent agent paradigm became widely accepted during the 1990s, and currently serves as the definition of the field.[a]

The paradigm has other advantages for AI. It provides a reliable and scientific way to test programs; researchers can directly compare or even combine different approaches to isolated problems, by asking which agent is best at maximizing a given "goal function". It also gives them a common language to communicate with other fields — such as mathematical optimization (which is defined in terms of "goals") or economics (which uses the same definition of a "rational agent").[166]

Evaluating approaches to AI

No established unifying theory or paradigm has guided AI research for most of its history.[s] The unprecedented success of statistical machine learning in the 2010s eclipsed all other approaches (so much so that some sources, especially in the business world, use the term "artificial intelligence" to mean "machine learning with neural networks"). This approach is mostly sub-symbolic, neat, soft and narrow (see below). Critics argue that these questions may have to be revisited by future generations of AI researchers.

Symbolic AI and its limits

Main articles: Symbolic AI, Physical symbol systems hypothesis, Moravec's paradox, and Dreyfus' critique of artificial intelligence

Symbolic AI (or "GOFAI")[168] simulated the high-level conscious reasoning that people use when they solve puzzles, express legal reasoning and do mathematics. They were highly successful at "intelligent" tasks such as algebra or IQ tests. In the 1960s, Newell and Simon proposed the physical symbol systems hypothesis: "A physical symbol system has the necessary and sufficient means of general intelligent action."[169]

However, the symbolic approach failed dismally on many tasks that humans solve easily, such as learning, recognizing an object or commonsense reasoning. Moravec's paradox is the discovery that high-level "intelligent" tasks were easy for AI, but low level "instinctive" tasks were extremely difficult.[170] Philosopher Hubert Dreyfus had argued since the 1960s that human expertise depends on unconscious instinct rather than conscious symbol manipulation, and on having a "feel" for the situation, rather than explicit symbolic knowledge.[171] Although his arguments had been ridiculed and ignored when they were first presented, eventually AI research came to agree.[t][48]

The issue is not resolved: sub-symbolic reasoning can make many of the same inscrutable mistakes that human intuition does, such as algorithmic bias. Critics such Noam Chomsky argue continuing research into symbolic AI will still be necessary to attain general intelligence,[173][174] in part because sub-symbolic AI is a move away from explainable AI: it can be difficult or impossible to understand why a modern statistical AI program made a particular decision.

Neat vs. scruffy

Main article: Neats and scruffies

"Neats" hope that intelligent behavior be described using simple, elegant principles (such as logic, optimization, or neural networks). "Scruffies" expect that it necessarily requires solving a large number of unrelated problems. This issue was actively discussed in the 70s and 80s,[175] but in the 1990s mathematical methods and solid scientific standards became the norm, a transition that Russell and Norvig termed "the victory of the neats".[176]

Soft vs. hard computing

Main article: Soft computing

Finding a provably correct or optimal solution is intractable for many important problems.[47] Soft computing is a set of techniques, including genetic algorithms, fuzzy logic and neural networks, that are tolerant of imprecision, uncertainty, partial truth and approximation. Soft computing was introduced in the late 80s and most successful AI programs in the 21st century are examples of soft computing with neural networks.

Narrow vs. general AI

Main article: Artificial general intelligence

AI researchers are divided as to whether to pursue the goals of artificial general intelligence and superintelligence (general AI) directly, or to solve as many specific problems as possible (narrow AI) in hopes these solutions will lead indirectly to the field's long-term goals[177][178] General intelligence is difficult to define and difficult to measure, and modern AI has had more verifiable successes by focussing on specific problems with specific solutions. The experimental sub-field of artificial general intelligence studies this area exclusively.

Machine consciousness, sentience and mind

Main articles: Philosophy of artificial intelligence and Artificial Consciousness

The philosophy of mind does not know whether a machine can have a mind, consciousness and mental states, in the same sense that human beings do. This issue considers the internal experiences of the machine, rather than its external behavior. Mainstream AI research considers this issue irrelevant, because it does not effect the goals of the field. Stuart Russell and Peter Norvig observe that most AI researchers "don't care about the [philosophy of AI] — as long as the program works, they don't care whether you call it a simulation of intelligence or real intelligence."[179] However, the question has become central to the philosophy of mind. It is also typically the central question at issue in artificial intelligence in fiction.

Consciousness

Main articles: Hard problem of consciousness and Theory of mind

David Chalmers identified two problems in understanding the mind, which he named the "hard" and "easy" problems of consciousness.[180] The easy problem is understanding how the brain processes signals, makes plans and controls behavior. The hard problem is explaining how this feels or why it should feel like anything at all. Human information processing is easy to explain, however human subjective experience is difficult to explain. For example, it is easy to imagine a color blind person who has learned to identify which objects in their field of view are red, but it is not clear what would be required for the person to know what red looks like.[181]

Computationalism and functionalism

Main articles: Computationalism, Functionalism (philosophy of mind), and Chinese room

Computationalism is the position in the philosophy of mind that the human mind is an information processing system and that thinking is a form of computing. Computationalism argues that the relationship between mind and body is similar or identical to the relationship between software and hardware and thus may be a solution to the mind-body problem. This philosophical position was inspired by the work of AI researchers and cognitive scientists in the 1960s and was originally proposed by philosophers Jerry Fodor and Hilary Putnam.[182]

Philosopher John Searle characterized this position as "strong AI": "The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds."[u] Searle counters this assertion with his Chinese room argument, which attempts to show that, even if a machine perfectly simulates human behavior, there is still no reason to suppose it also has a mind.[185]

Robot rights

Main article: Robot rights

If a machine has a mind and subjective experience, then it may also have sentience (the ability to feel), and if so, then it could also suffer, and thus it would be entitled to certain rights.[186] Any hypothetical robot rights would lie on a spectrum with animal rights and human rights.[187] This issue has been considered in fiction for centuries,[188] and is now being considered by, for example, California's Institute for the Future, however critics argue that the discussion is premature.[189]

Future

Superintelligence

Main articles: Superintelligence, Technological singularity, and Transhumanism

A superintelligence, hyperintelligence, or superhuman intelligence, is a hypothetical agent that would possess intelligence far surpassing that of the brightest and most gifted human mind. Superintelligence may also refer to the form or degree of intelligence possessed by such an agent.[178]

If research into artificial general intelligence produced sufficiently intelligent software, it might be able to reprogram and improve itself. The improved software would be even better at improving itself, leading to recursive self-improvement.[190] Its intelligence would increase exponentially in an intelligence explosion and could dramatically surpass humans. Science fiction writer Vernor Vinge named this scenario the "singularity".[191] Because it is difficult or impossible to know the limits of intelligence or the capabilities of superintelligent machines, the technological singularity is an occurrence beyond which events are unpredictable or even unfathomable.[192]

Robot designer Hans Moravec, cyberneticist Kevin Warwick, and inventor Ray Kurzweil have predicted that humans and machines will merge in the future into cyborgs that are more capable and powerful than either. This idea, called transhumanism, has roots in Aldous Huxley and Robert Ettinger.[193]

Edward Fredkin argues that "artificial intelligence is the next stage in evolution", an idea first proposed by Samuel Butler's "Darwin among the Machines" as far back as 1863, and expanded upon by George Dyson in his book of the same name in 1998.[194]

Risks

Technological unemployment

Main articles: Workplace impact of artificial intelligence and Technological unemployment

In the past technology has tended to increase rather than reduce total employment, but economists acknowledge that "we're in uncharted territory" with AI.[195] A survey of economists showed disagreement about whether the increasing use of robots and AI will cause a substantial increase in long-term unemployment, but they generally agree that it could be a net benefit, if productivity gains are redistributed.[196] Subjective estimates of the risk vary widely; for example, Michael Osborne and Carl Benedikt Frey estimate 47% of U.S. jobs are at "high risk" of potential automation, while an OECD report classifies only 9% of U.S. jobs as "high risk".[v][198]

Unlike previous waves of automation, many middle-class jobs may be eliminated by artificial intelligence; The Economist states that "the worry that AI could do to white-collar jobs what steam power did to blue-collar ones during the Industrial Revolution" is "worth taking seriously".[199] Jobs at extreme risk range from paralegals to fast food cooks, while job demand is likely to increase for care-related professions ranging from personal healthcare to the clergy.[200]

Bad actors and weaponized AI

Main articles: lethal autonomous weapon and artificial intelligence arms race

AI provides a number of tools that are particularly useful for authoritarian governments: smart spyware, face recognition and voice recognition allow widespread surveillance; such surveillance allows machine learning to classify potential enemies of the state and can prevent them from hiding; recommendation systems can precisely target propaganda and misinformation for maximum effect; deepfakes aid in producing misinformation; advanced AI can make centralized decision making more competitive with liberal and decentralized systems such as markets.[201]

Terrorists, criminals and rogue states may use other forms of weaponized AI such as advanced digital warfare and lethal autonomous weapons. By 2015, over fifty countries were reported to be researching battlefield robots.[202]

Algorithmic bias

Main article: algorithmic bias

AI programs can become biased after learning from real world data. It is not typically introduced by the system designers, but is learned by the program, and thus the programmers are often unaware that the bias exists.[203] Bias can be inadvertently introduced by the way training data is selected.[204] It can also emerge from correlations: AI is used to classify individuals into groups and then make predictions assuming that the individual will resemble other members of the group. In some cases, this assumption may be unfair.[205] An example of this is COMPAS, a commercial program widely used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. ProPublica claims that the COMPAS-assigned recidivism risk level of black defendants is far more likely to be an overestimate than that of white defendants, despite the fact that the program was not told the races of the defendants.[206] Other examples where algorithmic bias can lead to unfair outcomes are when AI is used for credit rating or hiring.

Existential risk

Main articles: Existential risk from artificial general intelligence and Superintelligence

Superintelligent AI may be able to improve itself to the point that humans could not control it. This could, as physicist Stephen Hawking puts it, "spell the end of the human race".[207] Philosopher Nick Bostrom argues that sufficiently intelligent AI, if it chooses actions based on achieving some goal, will exhibit convergent behavior such as acquiring resources or protecting itself from being shut down. If this AI's goals do not fully reflect humanity's, it might need to harm humanity in order to acquire more resources or prevent itself from being shut down, ultimately to better achieve its goal. He concludes that AI poses a risk to mankind, however humble or "friendly" its stated goals might be.[208] Political scientist Charles T. Rubin argues that "any sufficiently advanced benevolence may be indistinguishable from malevolence." Humans should not assume machines or robots would treat us favorably because there is no a priori reason to believe that they would share our system of morality.[209]

The opinion of experts and industry insiders is mixed, with sizable fractions both concerned and unconcerned by risk from eventual superhumanly-capable AI.[210] Stephen Hawking, Microsoft founder Bill Gates, history professor Yuval Noah Harari, and SpaceX founder Elon Musk have all expressed serious misgivings about the future of AI.[211] Prominent tech titans including Peter Thiel (Amazon Web Services) and Musk have committed more than $1 billion to nonprofit companies that champion responsible AI development, such as OpenAI and the Future of Life Institute.[212] Mark Zuckerberg (CEO, Facebook) has said that artificial intelligence is helpful in its current form and will continue to assist humans.[213] Other experts argue is that the risks are far enough in the future to not be worth researching, or that humans will be valuable from the perspective of a superintelligent machine.[214] Rodney Brooks, in particular, has said that "malevolent" AI is still centuries away.[w]

Ethical machines

Main articles: Machine ethics, Friendly AI, Artificial moral agents, and Human Compatible

Friendly AI are machines that have been designed from the beginning to minimize risks and to make choices that benefit humans. Eliezer Yudkowsky, who coined the term, argues that developing friendly AI should be a higher research priority: it may require a large investment and it must be completed before AI becomes an existential risk.[216]

Machines with intelligence have the potential to use their intelligence to make ethical decisions. The field of machine ethics provides machines with ethical principles and procedures for resolving ethical dilemmas.[217] Machine ethics is also called machine morality, computational ethics or computational morality,[217] and was founded at an AAAI symposium in 2005.[218]

Others approaches include Wendell Wallach's "artificial moral agents"[219] and Stuart J. Russell's three principles for developing provably beneficial machines.[220]