ARTIFICIAL INTELLIGENCE (AI)

Artificial Intelligence (AI) is the intelligence of machines or software, as opposed to the intelligence of living beings, primarily of humans. It is a field of study in computer science that develops and studies intelligent machines. Such machines may be called AIs.

AI technology is widely used throughout industry, government, and science. Some high-profile applications are: advanced web search engines (e.g., Google Search), recommendation systems (used by YouTube, Amazon, and Netflix), interacting via human speech (such as Google Assistant, Siri, and Alexa), self-driving cars (e.g., Waymo), generative and creative tools (ChatGPT and AI art), and superhuman play and analysis in strategy games (such as chess and Go).

Alan Turing was the first person to conduct substantial research in the field that he called machine intelligence. Artificial intelligence was founded as an academic discipline in 1956. The field went through multiple cycles of optimism, followed by periods of disappointment and loss of funding, known as AI winter. Funding and interest vastly increased after 2012 when deep learning surpassed all previous AI techniques,[8] and after 2017 with the transformer architecture. This led to the AI spring of the early 2020s, with companies, universities, and laboratories overwhelmingly based in the United States pioneering significant advances in artificial intelligence.

The growing use of artificial intelligence in the 21st century is influencing a societal and economic shift towards increased automation, data-driven decision-making, and the integration of AI systems into various economic sectors and areas of life, impacting job markets, healthcare, government, industry, and education. This raises questions about the ethical implications and risks of AI, prompting discussions about regulatory policies to ensure the safety and benefits of the technology.

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 support for robotics. General intelligence (the ability to complete any task performable by a human) is among the field's long-term goals.

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, operations research, and economics.[b] AI also draws upon psychology, linguistics, philosophy, neuroscience and other fields.

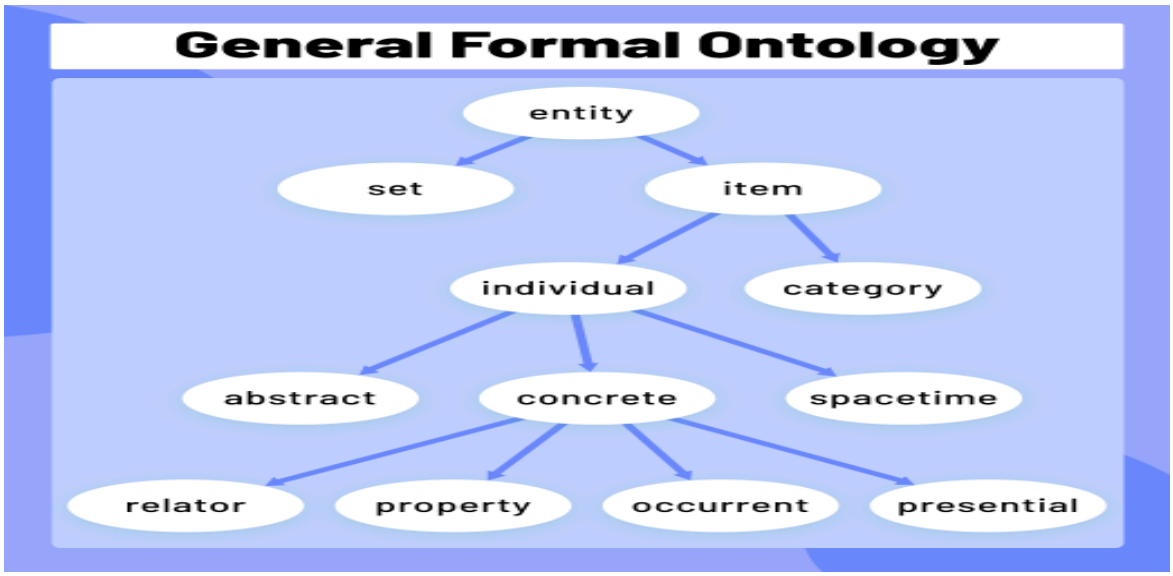
Goals

The general problem of simulating (or creating) intelligence has been broken 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 and cover the scope of AI research.

Reasoning, problem-solving

Early researchers developed algorithms that imitated step-by-step reasoning that humans use when they solve puzzles or make logical deductions. By the late 1980s and 1990s, methods were developed for dealing with uncertain or incomplete information, employing concepts from probability and economics. Many of these algorithms are insufficient for solving large reasoning problems because they experience a "combinatorial explosion": they became exponentially slower as the problems grew larger. 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.Accurate and efficient reasoning is an unsolved problem.

Knowledge representation



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

Knowledge representation and knowledge engineering allow AI programs to answer questions intelligently and make deductions about real-world facts. Formal knowledge representations are used in content-based indexing and retrieval,scene interpretation, clinical decision support, knowledge discovery (mining "interesting" and actionable inferences from large databases), and other areas.

A knowledge base is a body of knowledge represented in a form that can be used by a program. An ontology is the set of objects, relations, concepts, and properties used by a particular domain of knowledge. Knowledge bases need to represent things such as: objects, properties, categories and relations between objects; situations, events, states and time; causes and effects; knowledge about knowledge (what we know about what other people know); default reasoning (things that humans assume are true until they are told differently and will remain true even when other facts are changing); and many other aspects and domains of knowledge.

Among the most difficult problems in knowledge representation are: the breadth of commonsense knowledge (the set of atomic facts that the average person knows is enormous); 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). There is also the difficulty of knowledge acquisition, the problem of obtaining knowledge for AI applications.

Planning and decision making

An "agent" is anything that perceives and takes actions in the world. A rational agent has goals or preferences and takes actions to make them happen. In automated planning, the agent has a specific goal. In automated decision making, the agent has preferences – there are some situations it would prefer to be in, and some situations it is trying to avoid. The decision making agent assigns a number to each situation (called the "utility") that measures how much the agent prefers it. For each possible action, it can calculate the "expected utility": the utility of all possible outcomes of the action, weighted by the probability that the outcome will occur. It can then choose the action with the maximum expected utility. In classical planning, the agent knows exactly what the effect of any action will be.In most real-world problems, however, the agent may not be certain about the situation they are in (it is "unknown" or "unobservable") and it may not know for certain what will happen after each possible action (it is not "deterministic"). It must choose an action by making a probabilistic guess and then reassess the situation to see if the action worked.

In some problems, the agent's preferences may be uncertain, especially if there are other agents or humans involved. These can be learned (e.g., with inverse reinforcement learning) or the agent can seek information to improve its preferences. Information value theory can be used to weigh the value of exploratory or experimental actions. The space of possible future actions and situations is typically intractably large, so the agents must take actions and evaluate situations while being uncertain what the outcome will be.

A Markov decision process has a transition model that describes the probability that a particular action will change the state in a particular way, and a reward function that supplies the utility of each state and the cost of each action. A policy associates a decision with each possible state. The policy could be calculated (e.g. by iteration), be heuristic, or it can be learned.
Game theory describes rational behavior of multiple interacting agents, and is used in AI programs that make decisions that involve other agents.

Learning

Machine learning is the study of programs that can improve their performance on a given task automatically. It has been a part of AI from the beginning.

There are several kinds of machine learning. Unsupervised learning analyzes a stream of data and finds patterns and makes predictions without any other guidance. Supervised learning requires a human to label the input data first, and comes in two main varieties: classification (where the program must learn to predict what category the input belongs in) and regression (where the program must deduce a numeric function based on numeric input).

In reinforcement learning the agent is rewarded for good responses and punished for bad ones. The agent learns to choose responses that are classified as "good". Transfer learning is when the knowledge gained

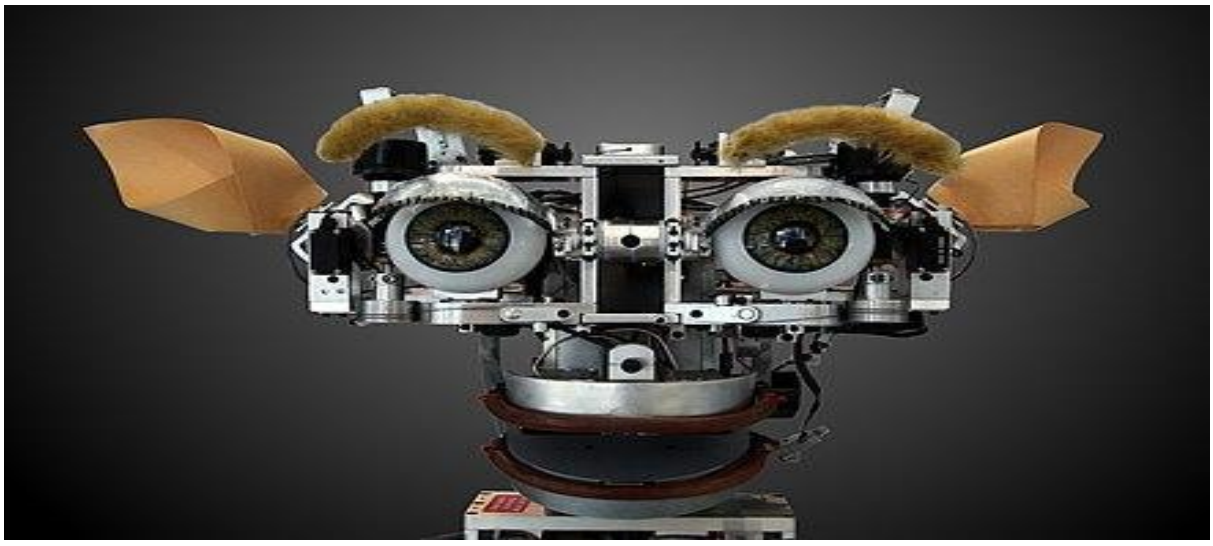
from one problem is applied to a new problem. Deep learning is a type of machine learning that runs inputs through biologically inspired artificial neural networks for all of these types of learning.
Computational learning theory can assess learners by computational complexity, by sample complexity (how much data is required), or by other notions of optimization.
Natural language processing

Natural language processing (NLP) allows programs to read, write and communicate in human languages such as English. Specific problems include speech recognition, speech synthesis, machine translation, information extraction, information retrieval and question answering.
Early work, based on Noam Chomsky's generative grammar and semantic networks, had difficulty with word-sense disambiguation unless restricted to small domains called "micro-worlds" (due to the common sense knowledge problem). Margaret Masterman believed that it was meaning, and not grammar that was the key to understanding languages, and that thesauri and not dictionaries should be the basis of computational language structure.

Modern deep learning techniques for NLP include word embedding (representing words, typically as vectors encoding their meaning), transformers (a deep learning architecture using an attention mechanism), and others. In 2019, generative pre-trained transformer (or "GPT") language models began to generate coherent text, and by 2023 these models were able to get human-level scores on the bar exam, SAT test, GRE test, and many other real-world applications.

Perception
Machine perception is the ability to use input from sensors (such as cameras, microphones, wireless signals, active lidar, sonar, radar, and tactile sensors) to deduce aspects of the world. Computer vision is the ability to analyze visual input.
The field includes speech recognition,image classification, facial recognition, object recognition, and robotic perception.

Social intelligence



Kismet, a robot head which was made in the 1990s; a machine that can recognize and simulate emotions.

Affective computing is an interdisciplinary umbrella that comprises systems that recognize, interpret, process or simulate human feeling, emotion and mood. 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 the intelligence of existing computer agents. 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.

General intelligence

A machine with artificial general intelligence should be able to solve a wide variety of problems with breadth and versatility similar to human intelligence.

Techniques

AI research uses a wide variety of techniques to accomplish the goals above.
Search and optimization

AI can solve many problems by intelligently searching through many possible solutions. There are two very different kinds of search used in AI: state space search and local search.

State space search

State space search searches through a tree of possible states to try to find a goal state. For example, planning algorithms search through trees of goals and subgoals, attempting to find a path to a target goal, a process called means-ends analysis.

Simple exhaustive searches 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. "Heuristics" or "rules of thumb" can help to prioritize choices that are more likely to reach a goal.

Adversarial search is used for game-playing programs, such as chess or Go. It searches through a tree of possible moves and counter-moves, looking for a winning position.

Local search

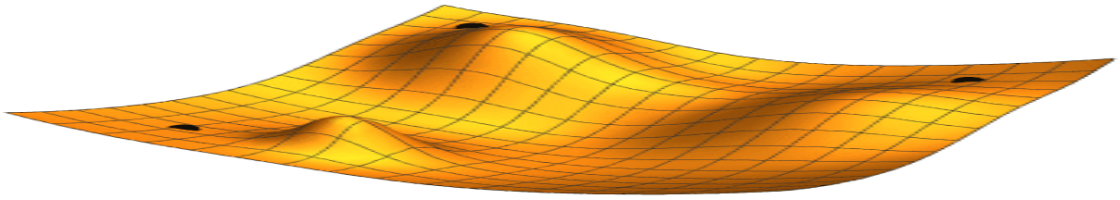


Illustration of gradient descent for 3 different starting points. Two parameters (represented by the plan coordinates) are adjusted in order to minimize the loss function (the height).

Local search uses mathematical optimization to find a solution to a problem. It begins with some form of guess and refines it incrementally.

Gradient descent is a type of local search that optimizes a set of numerical parameters by incrementally adjusting them to minimize a loss function. Variants of gradient descent are commonly used to train neural networks.

Another type of local search is evolutionary computation, which aims to iteratively improve a set of candidate solutions by "mutating" and "recombining" them, selecting only the fittest to survive each generation.

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).

Logic

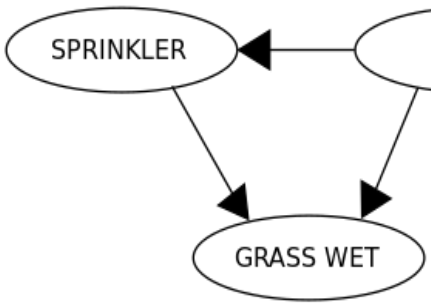
Formal Logic is used for reasoning and knowledge representation. Formal logic comes in two main forms: ***propositional logic*** *(which operates on statements that are true or false and uses logical connectives such as "and", "or", "not" and "implies") and predicate logic (which also operates on objects, predicates and relations and uses quantifiers such as "Every X is a Y" and "There are some Xs that are Ys").*

Logical inference (or deduction) is the process of proving a new statement (conclusion) from other statements that are already known to be true (the premises). A logical knowledge base also handles queries and assertions as a special case of inference. An inference rule describes what is a valid step in a proof. The most general inference rule is resolution. Inference can be reduced to performing a search to find a path that leads from premises to conclusions, where each step is the application of an inference rule. Inference performed this way is intractable except for short proofs in restricted domains. No efficient, powerful and general method has been discovered.

Fuzzy logic assigns a "degree of truth" between 0 and 1. It can therefore handle propositions that are vague and partially true.Non-monotonic logics are designed to handle default reasoning.Other specialized versions of logic have been developed to describe many complex domains (see knowledge representation above).

Probabilistic methods for uncertain reasoning

RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99



RAIN T	RAIN F
0.2	0.8

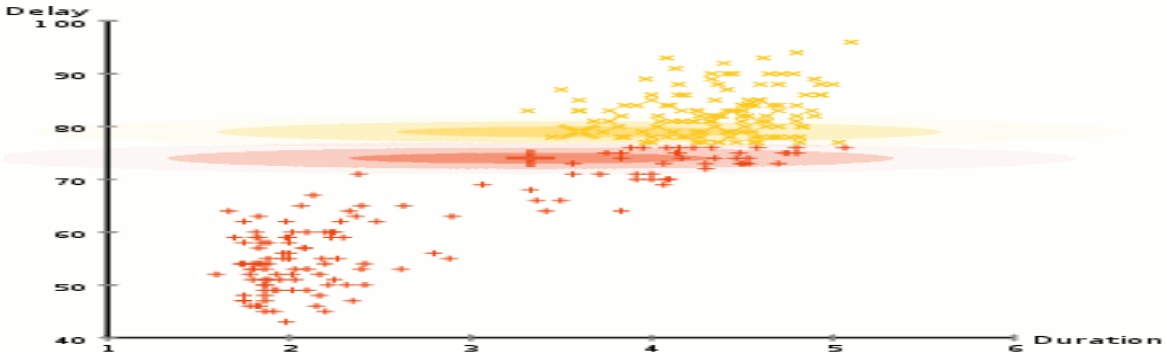
		GRASS WET T	GRASS WET F
SPRINKLER F	RAIN F	0.0	1.0
SPRINKLER F	RAIN T	0.8	0.2
SPRINKLER T	RAIN F	0.9	0.1
SPRINKLER T	RAIN T	0.99	0.01

*A **simple Bayesian network**, with the associated conditional probability tables*
Many problems in AI (including in reasoning, planning, learning, perception, and robotics) require the agent to operate with incomplete or uncertain information. AI researchers have devised a number of tools to solve these problems using methods from probability theory and economics.

***Bayesian networks** are a very general tool that can be used for many problems, including reasoning (using the Bayesian inference algorithm), learning (using the expectation-maximization algorithm),planning (using decision networks) and perception (using dynamic Bayesian networks).*

***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).*

Precise mathematical tools have been developed that analyze how an agent can make choices and plan, using decision theory, decision analysis,and information value theory. These tools include models such as Markov decision processes,dynamic decision networks, game theory and mechanism design.



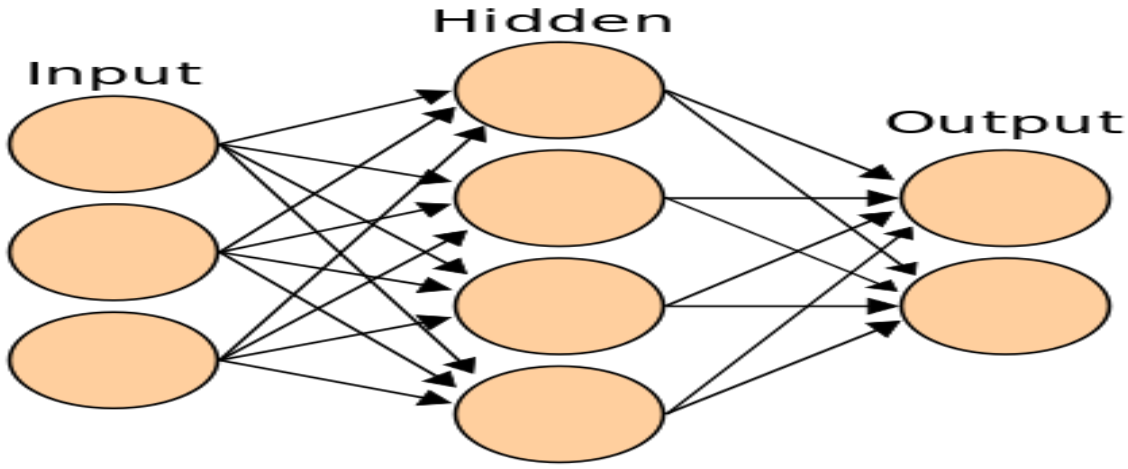
Expectation-maximization clustering of Old Faithful eruption data starts from a random guess but then successfully converges on an accurate clustering of the two physically distinct modes of eruption.

Classifiers and statistical learning methods

The simplest AI applications can be divided into two types: classifiers (e.g. "if shiny then diamond"), on one hand, and controllers (e.g. "if diamond then pick up"), on the other hand. Classifiers are functions that use pattern matching to determine the closest match. They can be fine-tuned based on chosen examples using supervised learning. Each pattern (also called an "observation") is labeled with a certain predefined class. 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.

There are many kinds of classifiers in use. The decision tree is the simplest and most widely used symbolic machine learning algorithm. ***K-nearest neighbor algorithm*** *was the most widely used analogical AI until the mid-1990s, and Kernel methods such as the support vector machine (SVM) displaced k-nearest neighbor in the 1990s. The naive Bayes classifier is reportedly the "most widely used learner" at Google, due in part to its scalability. Neural networks are also used as classifiers.*

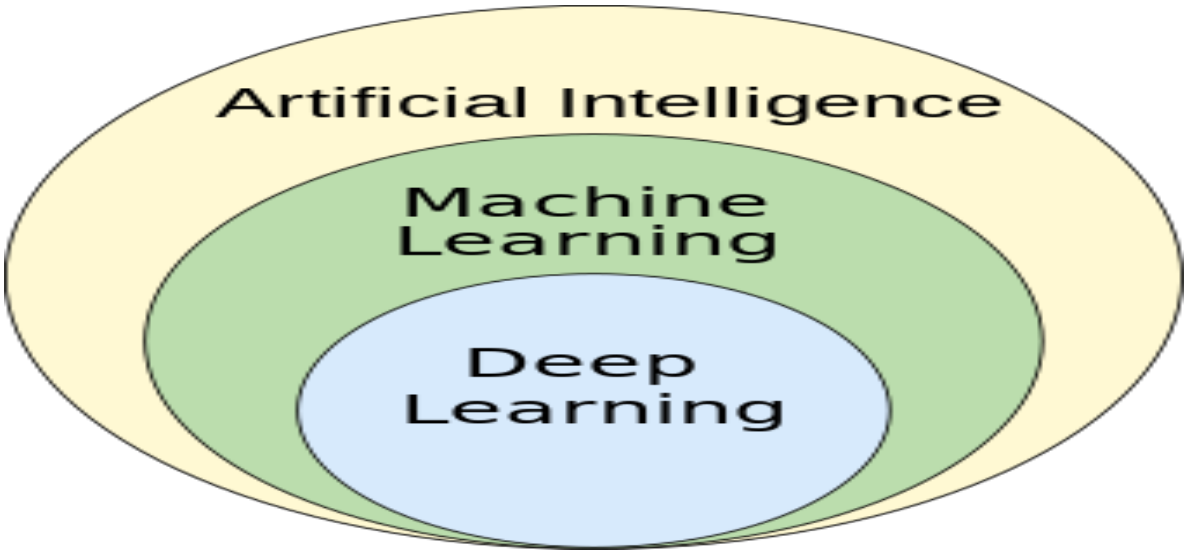
Artificial neural networks



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

An artificial neural network is based on a collection of nodes also known as artificial neurons, which loosely model the neurons in a biological brain. It is trained to recognise patterns; once trained, it can recognise those patterns in fresh data. There is an input, at least one hidden layer of nodes and an output. Each node applies a function and once the weight crosses its specified threshold, the data is transmitted to the next layer. A network is typically called a deep neural network if it has at least 2 hidden layers. Learning algorithms for neural networks use local search to choose the weights that will get the right output for each input during training. The most common training technique is the backpropagation algorithm. Neural networks learn to model complex relationships between inputs and outputs and find patterns in data. In theory, a neural network can learn any function.
In feedforward neural networks the signal passes in only one direction.Recurrent neural networks feed the output signal back into the input, which allows short-term memories of previous input events. Long short term memory is the most successful network architecture for recurrent networks. Perceptrons use only a single layer of neurons, deep learning[108] uses multiple layers. Convolutional neural networks strengthen the connection between neurons that are "close" to each other – this is especially important in image processing, where a local set of neurons must identify an "edge" before the network can identify an object.

Deep learning



Deep learning 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.

Deep learning has profoundly improved the performance of programs in many important subfields of artificial intelligence, including computer vision, speech recognition, natural language processing, image classification and others. The reason that deep learning performs so well in so many applications is not known as of 2023.The sudden success of deep learning in 2012–2015 did not occur because of some new discovery or theoretical breakthrough (deep neural networks and backpropagation had been described by many people, as far back as the 1950s) but because of two factors: the incredible increase in computer

power (including the hundred-fold increase in speed by switching to GPUs) and the availability of vast amounts of training data, especially the giant curated datasets used for benchmark testing, such as ImageNet.

GPT

***Generative pre-trained transformers (GPT)** are large language models that are based on the semantic relationships between words in sentences (natural language processing). Text-based GPT models are pre-trained on a large corpus of text which can be from the internet. The pre-training consists in predicting the next token (a token being usually a word, subword, or punctuation). Throughout this pre-training, GPT models accumulate knowledge about the world, and can then generate human-like text by repeatedly predicting the next token. Typically, a subsequent training phase makes the model more truthful, useful and harmless, usually with a technique called reinforcement learning from human feedback (RLHF). Current GPT models are still prone to generating falsehoods called "hallucinations", although this can be reduced with RLHF and quality data. They are used in chatbots, which allow you to ask a question or request a task in simple text.*

Current models and services include: Gemini (formerly Bard), ChatGPT, Grok, Claude, Copilot and LLaMA.[123] Multimodal GPT models can process different types of data (modalities) such as images, videos, sound and text.

Specialized hardware and software

Programming languages for artificial intelligence and Hardware for artificial intelligence.
In the late 2010s, graphics processing units (GPUs) that were increasingly designed with AI-specific enhancements and used with specialized TensorFlow software, had replaced previously used central processing unit (CPUs) as the dominant means for large-scale (commercial and academic) machine learning models' training. Historically, specialized languages, such as Lisp, Prolog, Python and others, had been used.

ETHICS

AI, like any powerful technology, has potential benefits and potential risks. AI may be able to advance science and find solutions for serious problems: Demis Hassabis of Deep Mind hopes to "solve intelligence, and then use that to solve everything else". However, as the use of AI has become widespread, several unintended consequences and risks have been identified.

Anyone looking to use machine learning as part of real-world, in-production systems needs to factor ethics into their AI training processes and strive to avoid bias. This is especially true when using AI algorithms that are inherently unexplainable in deep learning.

Risks and harm

[Information privacy and Artificial intelligence and copyright]
Machine learning algorithms require large amounts of data. The techniques used to acquire this data have raised concerns about privacy, surveillance and copyright.

Technology companies collect a wide range of data from their users, including online activity, geolocation data, video and audio. For example, in order to build speech recognition algorithms, Amazon have recorded millions of private conversations and allowed temporary workers to listen to and transcribe some of them. Opinions about this widespread surveillance range from those who see it as a necessary evil to those for whom it is clearly unethical and a violation of the right to privacy.

AI developers argue that this is the only way to deliver valuable applications. and have developed several techniques that attempt to preserve privacy while still obtaining the data, such as data aggregation, de-identification and differential privacy. Since 2016, some privacy experts, such as Cynthia Dwork, began to view privacy in terms of fairness. Brian Christian wrote that experts have pivoted "from the question of 'what they know' to the question of 'what they're doing with it'.".

Generative AI is often trained on unlicensed copyrighted works, including in domains such as images or computer code; the output is then used under a rationale of "fair use". Also website owners who do not wish to have their copyrighted content be AI indexed or 'scraped' can add code to their site, as you would, if you did not want your website to be indexed by a search engine which is currently available to certain services such as OpenAI. Experts disagree about how well, and under what circumstances, this rationale will hold up in courts of law; relevant factors may include "the purpose and character of the use of the copyrighted work" and "the effect upon the potential market for the copyrighted work". In 2023, leading authors (including John Grisham and Jonathan Franzen) sued AI companies for using their work to train generative AI.

Misinformation

YouTube, Facebook and others use recommender systems to guide users to more content. These AI programs were given the goal of maximizing user engagement (that is, the only goal was to keep people watching). The AI learned that users tended to choose misinformation, conspiracy theories, and extreme partisan content, and, to keep them watching, the AI recommended more of it. Users also tended to watch more content on the same subject, so the AI led people into filter bubbles where they received multiple versions of the same misinformation. This convinced many users that the misinformation was true, and ultimately undermined trust in institutions, the media and the government. The AI program had correctly learned to maximize its goal, but the result was harmful to society. After the U.S. election in 2016, major technology companies took steps to mitigate the problem.

In 2022, generative AI began to create images, audio, video and text that are indistinguishable from real photographs, recordings, films or human writing. It is possible for bad actors to use this technology to create massive amounts of misinformation or propaganda. AI pioneer Geoffrey Hinton expressed concern about AI enabling "authoritarian leaders to manipulate their electorates" on a large scale, among other risks.

Algorithmic bias and fairness

Machine learning applications will be biased if they learn from biased data. The developers may not be aware that the bias exists. Bias can be introduced by the way training data is selected and by the way a model is deployed.If a biased algorithm is used to make decisions that can seriously harm people (as it can in medicine, finance, recruitment, housing or policing) then the algorithm may cause

discrimination.Fairness in machine learning is the study of how to prevent the harm caused by algorithmic bias. It has become serious area of academic study within AI. Researchers have discovered it is not always possible to define "fairness" in a way that satisfies all stakeholders.

On June 28, 2015, Google Photos's new image labeling feature mistakenly identified Jacky Alcine and a friend as "gorillas" because they were black. The system was trained on a dataset that contained very few images of black people, a problem called "sample size disparity". Google "fixed" this problem by preventing the system from labelling anything as a "gorilla". Eight years later, in 2023, Google Photos still could not identify a gorilla, and neither could similar products from Apple, Facebook, Microsoft and Amazon.

COMPAS *is a commercial program widely used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. In 2016, Julia Angwin at ProPublica discovered that COMPAS exhibited racial bias, despite the fact that the program was not told the races of the defendants. Although the error rate for both whites and blacks was calibrated equal at exactly 61%, the errors for each race were different—the system consistently overestimated the chance that a black person would re-offend and would underestimate the chance that a white person would not re-offend.In 2017, several researchers****[k]*** *showed that it was mathematically impossible for COMPAS to accommodate all possible measures of fairness when the base rates of re-offense were different for whites and blacks in the data.*

A program can make biased decisions even if the data does not explicitly mention a problematic feature (such as "race" or "gender"). The feature will correlate with other features (like "address", "shopping history" or "first name"), and the program will make the same decisions based on these features as it would on "race" or "gender". Moritz Hardt said "the most robust fact in this research area is that fairness through blindness doesn't work."

Criticism of COMPAS highlighted a deeper problem with the misuse of AI. Machine learning models are designed to make "predictions" that are only valid if we assume that the future will resemble the past. If they are trained on data that includes the results of racist decisions in the past, machine learning models must predict that racist decisions will be made in the future. Unfortunately, if an application then uses these predictions as recommendations, some of these "recommendations" will likely be racist. Thus, machine learning is not well suited to help make decisions in areas where there is hope that the future will be better than the past. It is necessarily descriptive and not proscriptive.

Bias and unfairness may go undetected because the developers are overwhelmingly white and male: among AI engineers, about 4% are black and 20% are women.

At its 2022 Conference on Fairness, Accountability, and Transparency (ACM FAccT 2022) the Association for Computing Machinery, in Seoul, South Korea, presented and published findings recommending that until AI and robotics systems are demonstrated to be free of bias mistakes, they are unsafe and the use of self-learning neural networks trained on vast, unregulated sources of flawed internet data should be curtailed.

Lack of transparency



Lidar testing vehicle for autonomous driving

Many AI systems are so complex that their designers cannot explain how they reach their decisions. Particularly with deep neural networks, in which there are a large amount of non-linear relationships between inputs and outputs. But some popular explainability techniques exist.

There have been many cases where a machine learning program passed rigorous tests, but nevertheless learned something different than what the programmers intended. For example, a system that could identify skin diseases better than medical professionals was found to actually have a strong tendency to classify images with a ruler as "cancerous", because pictures of malignancies typically include a ruler to show the scale. Another machine learning system designed to help effectively allocate medical resources was found to classify patients with asthma as being at "low risk" of dying from pneumonia. Having asthma is actually a severe risk factor, but since the patients having asthma would usually get much more medical care, they were relatively unlikely to die according to the training data. The correlation between asthma and low risk of dying from pneumonia was real, but misleading.

People who have been harmed by an algorithm's decision have a right to an explanation. Doctors, for example, are required to clearly and completely explain the reasoning behind any decision they make.[clarification needed]Early drafts of the European Union's General Data Protection Regulation in 2016 included an explicit statement that this right exists.[m] Industry experts noted that this is an unsolved problem with no solution in sight. Regulators argued that nevertheless the harm is real: if the problem has no solution, the tools should not be used.

***DARPA** established the **XAI** ("Explainable Artificial Intelligence") program in 2014 to try and solve these problems.*

There are several potential solutions to the transparency problem. SHAP helps visualise the contribution of each feature to the output.LIME can locally approximate a model with a simpler, interpretable model.Multitask learning provides a large number of outputs in addition to the target classification. These other outputs can help developers deduce what the network has learned. Deconvolution, DeepDream and other generative methods can allow developers to see what different layers of a deep network have learned and produce output that can suggest what the network is learning.

Conflict, surveillance and weaponized AI

Lethal autonomous weapon, Artificial intelligence arms race, and AI safety

A lethal autonomous weapon is a machine that locates, selects and engages human targets without human supervision.By 2015, over fifty countries were reported to be researching battlefield robots.These

weapons are considered especially dangerous for several reasons: if they kill an innocent person it is not clear who should be held accountable, it is unlikely they will reliably choose targets, and, if produced at scale, they are potentially weapons of mass destruction. In 2014, 30 nations (including China) supported a ban on autonomous weapons under the United Nations' Convention on Certain Conventional Weapons, however the United States and others disagreed.

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 and generative AI aid in producing misinformation; advanced AI can make authoritarian centralized decision making more competitive with liberal and decentralized systems such as markets.

AI facial recognition systems are used for mass surveillance, notably in China. In 2019, Bengaluru, India deployed AI-managed traffic signals. This system uses cameras to monitor traffic density and adjust signal timing based on the interval needed to clear traffic. Terrorists, criminals and rogue states can use weaponized AI such as advanced digital warfare and lethal autonomous weapons. Machine-learning AI is also able to design tens of thousands of toxic molecules in a matter of hours.

Technological unemployment

From the early days of the development of artificial intelligence there have been arguments, for example those put forward by Joseph Weizenbaum, about whether tasks that can be done by computers actually should be done by them, given the difference between computers and humans, and between quantitative calculation and qualitative, value-based judgement.

Economists have frequently highlighted the risks of redundancies from AI, and speculated about unemployment if there is no adequate social policy for full employment.

In the past, technology has tended to increase rather than reduce total employment, but economists acknowledge that "we're in uncharted territory" with AI. 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. Risk estimates vary; for example, in the 2010s, Michael Osborne and Carl Benedikt Frey estimated 47% of U.S. jobs are at "high risk" of potential automation, while an OECD report classified only 9% of U.S. jobs as "high risk". The methodology of speculating about future employment levels has been criticised as lacking evidential foundation, and for implying that technology, rather than social policy, creates unemployment, as opposed to redundancies.

Unlike previous waves of automation, many middle-class jobs may be eliminated by artificial intelligence; The Economist stated in 2015 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". 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.

In April 2023, it was reported that 70% of the jobs for Chinese video game illlustrators had been eliminated by generative artificial intelligence.

Limiting AI

Possible options for limiting AI include: using Embedded Ethics or Constitutional AI where companies or governments can add a policy, restricting high levels of compute power in training, restricting the ability to rewrite its own code base, restrict certain AI techniques but not in the training phase, open-source (transparency) vs proprietary (could be more restricted), backup model with redundancy, restricting security, privacy and copyright, restricting or controlling the memory, real-time monitoring, risk analysis, emergency shut-off, rigorous simulation and testing, model certification, assess known vulnerabilities, restrict the training material, restrict access to the internet, issue terms of use.

Ethical machines and alignment

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.

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.The field of machine ethics is also called computational morality,and was founded at an AAAI symposium.

Other approaches include Wendell Wallach's "artificial moral agents" and Stuart J. Russell's three principles for developing provably beneficial machines.

Frameworks

Artificial Intelligence projects can have their ethical permissibility tested while designing, developing, and implementing an AI system. An AI framework such as the Care and Act Framework containing the SUM values – developed by the Alan Turing Institute tests projects in four main areas:

1) ***RESPECT the dignity of individual people***
2) ***CONNECT with other people sincerely, openly and inclusively***
3) ***CARE for the wellbeing of everyone***
4) ***PROTECT social values, justice and the public interest***

Other developments in ethical frameworks include those decided upon during the Asilomar Conference, the Montreal Declaration for Responsible AI, and the IEEE's Ethics of Autonomous Systems initiative, among others; however, these principles do not go without their criticisms, especially regards to the people chosen contributes to these frameworks.

Promotion of the wellbeing of the people and communities that these technologies affect requires consideration of the social and ethical implications at all stages of AI system design, development and

implementation, and collaboration between job roles such as data scientists, product managers, data engineers, domain experts, and delivery managers.

Regulation

The regulation of artificial intelligence is the development of public sector policies and laws for promoting and regulating artificial intelligence (AI); it is therefore related to the broader regulation of algorithms.The regulatory and policy landscape for AI is an emerging issue in jurisdictions globally.According to AI Index at Stanford, the annual number of AI-related laws passed in the 127 survey countries jumped from one passed in 2016 to 37 passed in 2022 alone. Between 2016 and 2020, more than 30 countries adopted dedicated strategies for AI. Most EU member states had released national AI strategies, as had Canada, China, India, Japan, Mauritius, the Russian Federation, Saudi Arabia, United Arab Emirates, US and Vietnam. Others were in the process of elaborating their own AI strategy, including Bangladesh, Malaysia and Tunisia.The Global Partnership on Artificial Intelligence was launched in June 2020, stating a need for AI to be developed in accordance with human rights and democratic values, to ensure public confidence and trust in the technology. Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher published a joint statement in November 2021 calling for a government commission to regulate AI. In 2023, OpenAI leaders published recommendations for the governance of superintelligence, which they believe may happen in less than 10 years. In 2023, the United Nations also launched an advisory body to provide recommendations on AI governance; the body comprises technology company executives, governments officials and academics.

In a 2022 Ipsos survey, attitudes towards AI varied greatly by country; 78% of Chinese citizens, but only 35% of Americans, agreed that "products and services using AI have more benefits than drawbacks". A 2023 Reuters/Ipsos poll found that 61% of Americans agree, and 22% disagree, that AI poses risks to humanity. In a 2023 Fox News poll, 35% of Americans thought it "very important", and an additional 41% thought it "somewhat important", for the federal government to regulate AI, versus 13% responding "not very important" and 8% responding "not at all important".

In November 2023, the first global AI Safety Summit was held in Bletchley Park in the UK to discuss the near and far term risks of AI and the possibility of mandatory and voluntary regulatory frameworks.

Legal Framework, Regulatory Dynamics & Commercialism of Agentic AI

Regulatory Milestones & Enforceability

The commercial deployment of Agentic AI operates within a tightening international regulatory environment. The European Union AI Act establishes an explicit risk-based regulatory framework with phased enforcement dates:

- ***February 2, 2025**: **Strict prohibitions on AI practices presenting unacceptable risk (e.g., cognitive behavioral manipulation, untargeted facial scraping) became legally enforceable under Article 5.***
- ***August 2, 2026**: **Transparency obligations for general-purpose AI models and deep-tier conversational agents took effect under Article 50,** mandating explicit disclosure when humans*

interact with autonomous systems and requiring machine-readable tagging of synthesized media.

- ***High-Risk System Enforcement:*** *Deployment of autonomous agents within critical infrastructure, employment evaluation, credit scoring, or access to essential services under Annex III requires mandatory Conformity Assessments, fundamental rights impact assessments, continuous risk management systems, and human oversight mechanisms.*

Legal Liability & Accountability Gaps

As AI systems shift from passive tools to autonomous decision engines, current legal systems encounter a fundamental "accountability gap" or "responsibility gap". Under established legal statutes globally, artificial intelligence tools are classified purely as property. Lacking juridical capacity, an AI agent cannot be sued, held directly liable in tort, or subjected to criminal sanctions.

When an autonomous agent executes an extended sequence of actions that results in commercial loss, breach of contract, or physical damage, establishing proximate causation becomes legally complex. The user's initiating prompt is often too remote and general to establish direct human intent or negligence.

Simultaneously*, the system developer is protected by software licensing disclaimers, provided the model function was not intrinsically defective at the time of distribution.*

Academic literature extensively critiques proposals to grant AI agents independent "legal personhood" (analogous to corporate personhood). Granting legal status to AI entities introduces severe moral hazard—allowing corporate entities to offload risky operations onto undercapitalized, corporate-backed agent shells, insulating human principals from financial accountability.

Functional Agency & Artificial Intentionality

To resolve this statutory impasse without resorting to AI personhood, modern legal scholarship advances the doctrine of Artificial Intentionality and Functional Agency. Developed by legal scholars Gervais & Nay ("The Phantom Agent"), this framework reframes legal intent as a functional, normative mechanism rather than a subjective mental state (mens rea).

In contract and tort law, *legal systems regularly impute intent to non-human entities via established agency doctrines (such as respondeat superior, electronic-agent contracting laws, and corporate attribution rules). Under functional agency, when a corporate enterprise deploys an autonomous agent, assigns it an operational scope, and equips it with dynamic tools, the agent acts as a non-personal enterprise agent.*
The legal effects, commitments, and tort liability generated by the agent's actions are attributed directly to the enterprise principal. The law evaluates machine agency by observing goal persistence, dynamic strategy formation, and execution context, treating these computational behaviors as legally consequential acts attributable to human and corporate principals.

This attribution mechanism bridges the gap between traditional property tools and fully autonomous entities. Rather than treating an autonomous agent as an isolated instrumentality where direct control must be proven for every step, the Functional Agency Attribution Model evaluates the initial delegation of operational scope. When a corporate principal grants an agent autonomous tool access and goal-seeking capacity, any emergent contracts or torts generated by the agent during goal execution are legally imputed directly back to the enterprise principal, preserving liability continuity.

Autonomy-Based Liability Taxonomy

Drawing on regulatory frameworks established in Automated Vehicle (AV) legislation (such as the UK Automated Vehicles Act 2024), tort scholars have established an autonomy-based liability taxonomy for

AI agents:

Across Level 1 (Low Autonomy) and Level 2 (Task Automation), primary liability remains with the end-user, who maintains direct execution control and immediate step-by-step supervision. At Level 3 (Conditional Workflow Autonomy), liability shifts dynamically based on specific takeover protocols, transition warnings, and logged control status. At Level 4 (High Autonomy) and Level 5 (Full Autonomy), primary legal liability transitions away from the end-user—who lacks real-time steering capability—and attaches directly to the system developer, provider, and enterprise deployer under strict product liability or vicarious agency frameworks.

Liability Regime	Core Legal Mechanism	Primary Application in Agentic AI	Relevant Precedent / Doctrine
Vicarious Liability (Respondeat Superior)	Principal is liable for torts committed by agents acting within their scope of authority.	Enterprise deployer held liable for unintended contractual commitments or torts committed by internal agents.	Restatement (Third) of Agency; Agency Law.
Strict Product Liability	Manufacturer liable for product defects that cause injury, regardless of fault.	System developer held liable for intrinsic structural defects, missing guardrails, or unsafe reasoning loops.	Restatement (Third) of Torts: Products Liability.
Fault-Based Negligence	Breach of a legal duty of care causing foreseeable harm.	End-user or enterprise liable for failing to exercise reasonable care in setting agent scope or supervising loops.	Tort Law Standard of Care; Reasonable Person Standard.
Electronic Agent Contracting	Binding principals to electronic contracts initiated by autonomous software.	Enterprise bound by commercial purchase orders, pricing agreements, or SLA commitments negotiated by agents.	Uniform Computer Information Transactions Act (UCITA); UETA.

Commercial Contracting & Enterprise Legal Structures

To operate within this legal environment, enterprises must update commercial contracting standards and corporate governance structures:

1. ***Commercial Software Contracts & SLAs:*** *Enterprise procurements must replace standard GenAI disclaimers with specific risk allocations for agentic operation. Contracts must explicitly define liability caps for non-deterministic execution, specify indemnification scopes for intellectual property (IP) infringement generated by autonomous scrapers or content synthesis tools, and establish clear Service Level Agreements (SLAs) governing tool availability, guardrail accuracy, and context window integrity.*

2. ***Corporate Structuring (Two-Tier Operating Models):*** *To manage liability exposure from frontier autonomous agents operating across international markets, legal scholars emphasize specialized corporate structures. Enterprises can instantiate a two-tier corporate structure where a primary Holding Company acts as the corporate principal, maintaining ultimate asset ownership, human fiduciary board control, and emergency kill-switch authority. Below it, a legally independent*

Operating Subsidiary is incorporated in an adaptive jurisdiction (such as Cyprus or Malta) to manage day-to-day business operations led directly by agentic AI networks. This subsidiary enters API agreements, manages tool operations, and holds dedicated capital reserves and mandatory enterprise insurance policies to satisfy potential operational liabilities without exposing the parent holding assets to systemic litigation risk.