In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and animals. Leading AI textbooks define the field as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving".

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. A quip in Tesler's Theorem says "AI is whatever hasn't been done yet." For instance, optical character recognition is frequently excluded from things considered to be AI , having become a routine technology. Modern machine capabilities generally classified as AI include successfully understanding human speech, competing at the highest level in strategic game systems (such as chess and Go), autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

Artificial intelligence was founded as an academic discipline in 1955, and in the years since has experienced several waves of optimism, followed by disappointment and the loss of funding (known as an "AI winter"), followed by new approaches, success and renewed funding. For most of its history, AI research has been divided into subfields that often fail to communicate with each other.These sub-fields are based on technical considerations, such as particular goals (e.g. "robotics" or "machine learning"), the use of particular tools ("logic" or artificial neural networks), or deep philosophical differences. Subfields have also been based on social factors (particular institutions or the work of particular researchers).

The traditional problems (or goals) of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception and the ability to move and manipulate objects. General intelligence is among the field's long-term goals. Approaches include statistical methods, computational intelligence, and traditional symbolic AI. Many tools are used in AI, including versions of search and mathematical optimization, artificial neural networks, and methods based on statistics, probability and economics. The AI field draws upon computer science, information engineering, mathematics, psychology, linguistics, philosophy, and many other fields.

A typical AI analyses its environment and takes actions that maximize its chance of success. An AI's intended utility function (or goal) can be simple ("1 if the AI wins a game of Go, 0 otherwise") or complex ("Do mathematically similar actions to the ones succeeded in the past"). Goals can be explicitly defined or induced. If the AI is programmed for "reinforcement learning", goals can be implicitly induced by rewarding some types of behaviour or punishing others. Alternatively, an evolutionary system can induce goals by using a "fitness function" to mutate and preferentially replicate high-scoring AI systems, similar to how animals evolved to innately desire certain goals such as finding food. Some AI systems, such as nearest-neighbour, instead of reason by analogy, these systems are not generally given goals, except to the degree that goals are implicit in their training data. Such systems can still be benchmarked if the non-goal system is framed as a system whose "goal" is to successfully accomplish its narrow classification task.

Many AI algorithms are capable of learning from data; they can enhance themselves by learning new heuristics (strategies, or "rules of thumb", that have worked well in the past), or can themselves write other algorithms. Some of the "learners" described below, including Bayesian networks, decision trees, and nearest-neighbor, could theoretically, (given infinite data, time, and memory) learn to approximate any function, including which combination of mathematical functions would best describe the world[citation needed]. These learners could therefore, derive all possible knowledge, by considering every possible hypothesis and matching them against the data. In practice, it is almost never possible to consider every possibility, because of the phenomenon of "combinatorial explosion", where the amount of time needed to solve a problem grows exponentially. Much of AI research involves figuring out how to identify and avoid considering broad range of possibilities that are unlikely to be beneficial. For example, when viewing a map and looking for the shortest driving route from Denver to New York in the East, one can in most cases skip looking at any path through San Francisco or other areas far to the West; thus, an AI wielding a pathfinding algorithm like A\* can avoid the combinatorial explosion that would ensue if every possible route had to be ponderously considered in turn.

The earliest (and easiest to understand) approach to AI was symbolism (such as formal logic): "If an otherwise healthy adult has a fever, then they may have influenza". A second, more general, approach is Bayesian inference: "If the current patient has a fever, adjust the probability they have influenza in such-and-such way". The third major approach, extremely popular in routine business AI applications, are analogizers such as SVM and nearest-neighbour: "After examining the records of known past patients whose temperature, symptoms, age, and other factors mostly match the current patient, X% of those patients turned out to have influenza". A fourth approach is harder to understand, and is inspired by how the brain's machinery works: the artificial neural network approach uses artificial "neurons" that can learn by comparing itself to the desired output and altering the strengths of the connections between its internal neurons to "reinforce" connections that seemed to be useful. These four main approaches can overlap with each other and with evolutionary systems; for example, neural nets can learn to make inferences, to generalize, and to make analogies. Some systems implicitly or explicitly use multiple of these approaches, alongside many other AI and non-AI algorithms; the best approach is often dependent on the problem.

Learning algorithms work on basis that strategies, algorithms, and inferences that worked well in the past are expected to continue working well in the future. These inferences can be obvious, such as "since the sun rose every morning for the last 10,000 days, it will probably rise tomorrow morning as well". They can be nuanced, such as "X% of families have geographically separate species with color variants, so there is a Y% chance that undiscovered black swans exist". Learners also work on the basis of "Occam's razor": The simplest theory that explains the data is the likeliest. Therefore, according to Occam's razor principle, a learner must be designed such that it favours simpler theories to complex theories, except in cases where the complex theory is proven significantly better.

Settling on a bad, overly complex theory gerrymandered to fit all the past training data is known as overfitting. Many systems attempt to reduce overfitting by rewarding a theory in accordance with how well it fits the data, but penalizing the theory in accordance with how complex the theory is. Besides classic overfitting, learners can also disappoint by "learning the wrong lesson". A toy example is that an image classifier trained only on pictures of brown horses and black cats might conclude that all brown patches are likely to be horses.A real-world example is that, unlike humans, current image classifiers don't determine the spatial relationship between components of the picture; instead, they learn abstract patterns of pixels that humans are oblivious to, but that linearly correlate with images of certain types of real objects. Faintly superimposing such a pattern on a legitimate image results in an "adversarial" image that the system misclassifies.

A self-driving car system may use a neural network to determine which parts of the picture seem to match previous training images of pedestrians, and then model those areas as slow-moving but somewhat unpredictable rectangular prisms that must be avoided. Compared with humans, existing AI lacks several features of human "commonsense reasoning"; most notably, humans have powerful mechanisms for reasoning about "naïve physics" such as space, time, and physical interactions. This enables even young children to easily make inferences like "If I roll this pen off a table, it will fall on the floor". Humans also have a powerful mechanism of "folk psychology" that helps them to interpret natural-language sentences such as "The city councilmen refused the demonstrators a permit because they advocated violence". (A generic AI has difficulty discerning whether the ones alleged to be advocating violence are the councilmen or the demonstrators.)[78][79][80] This lack of "common knowledge" means that AI often makes different mistakes than humans make, in ways that can seem incomprehensible. For example, existing self-driving cars cannot reason about the location nor the intentions of pedestrians in the exact way that humans do, and instead must use non-human modes of reasoning to avoid accidents