Stanford University
Human-Centered
Artificial Intelligence

Artificial Intelligence Definitions

Intelligence might be defined as the ability to learn and perform a range of techniques to solve problems and achieve goals—techniques that are appropriate to the context in an uncertain, ever-varying world. A fully pre-programmed factory robot is flexible, accurate, and consistent, but not intelligent.

Artificial intelligence (AI), is a term coined in 1955 by John McCarthy, Stanford's first faculty member in AI, who defined it as "the science and engineering of making intelligent machines." Much research has human program software agents with the knowledge to behave in a particular way, like playing chess, but today, we emphasize agents that can learn, just as human beings navigating our changing world.

Autonomous systems can independently plan and decide sequences of steps to achieve a specified goal without being micromanaged. A hospital delivery robot must autonomously navigate busy corridors to succeed in its task. In AI, autonomy doesn't have the sense of being self-governing common in politics or biology.

Machine Learning (ML) is the part of AI that studies how computer systems can improve their perception, knowledge, decisions, or actions based on experience or data. For this, ML draws from computer science, statistics, psychology, neuroscience, economics, and control theory.

In **supervised learning**, a computer learns to predict human-given labels, such as particular dog breeds based on labeled dog pictures. **Unsupervised learning** does not require labels, but sometimes adopts **self-supervised learning**, constructing its own prediction tasks such as trying to predict each successive word in a sentence. **Reinforcement learning** enables autonomy by allowing an agent to learn action sequences that optimize its total rewards, such as winning games, without explicit examples of good techniques.

Deep learning is the use of large multi-layer **(artificial) neural networks** that compute with continuous (real number) representations, similar to the hierarchically organized neurons in human brains. It is successfully employed for all types of ML, with better generalization from small data and better scaling to big data and compute budgets. A recent breakthrough is the **transformer**, a neural net architecture which flexibly incorporates context via an attention mechanism, allowing powerful and computationally efficient analysis and generation of sequences, such as words in a paragraph.

Foundation models are an emerging class of models, often transformers trained by self-supervision on large-scale broad data, that can be easily adapted to perform a wide range of downstream tasks. The best-known examples are large pretrained language models like GPT-3, but the term extends to models for all modalities of data and knowledge.

An **algorithm** is a precise list of steps to take, such as a computer program. AI systems contain algorithms, but typically just for a few parts like a learning or reward calculation method. Much of their behavior emerges via learning from data or experience, a fundamental shift in system design that Stanford alumnus Andrej Karpathy dubbed **software 2.0**.

Narrow AI is intelligent systems for particular tasks, e.g., **speech** or **facial recognition**. **Human-level AI**, or **artificial general intelligence (AGI)**, seeks broadly intelligent, context-aware machines. It is needed for effective, adaptable **social chatbots** or **human-robot interaction**.

Human-centered artificial intelligence is AI that seeks to augment the abilities of, address the societal needs of, and draw inspiration from human beings. It researches and builds effective partners and tools for people, such as a robot helper and companion for the elderly.

Text by Professor Christopher Manning, v 1.2, April 2022

Learn more at hai.stanford.edu