Artificial intelligence, or AI, is technology that enables computers and machines to simulate human intelligence and problem-solving capabilities.

On its own or combined with other technologies (e.g., sensors, geolocation, robotics) AI can perform tasks that would otherwise require human intelligence or intervention. Digital assistants, GPS guidance, autonomous vehicles, and generative AI tools (like Open AI's Chat GPT) are just a few examples of AI in the daily news and our daily lives.

As a field of computer science, artificial intelligence encompasses (and is often mentioned together with) machine learning and deep learning. These disciplines involve the development of AI algorithms, modeled after the decision-making processes of the human brain, that can 'learn' from available data and make increasingly more accurate classifications or predictions over time.

Artificial intelligence has gone through many cycles of hype, but even to skeptics, the release of ChatGPT seems to mark a turning point. The last time generative AI loomed this large, the breakthroughs were in computer vision, but now the leap forward is in natural language processing (NLP). Today, generative AI can learn and synthesize not just human language but other data types including images, video, software code, and even molecular structures.

Applications for AI are growing every day. But as the hype around the use of AI tools in business takes off, conversations around ai ethics and responsible ai become critically important. For more on where IBM stands on these issues, please read Building trust in AI.

Weak AI—also known as narrow AI or artificial narrow intelligence (ANI)—is AI trained and focused to perform specific tasks. Weak AI drives most of the AI that surrounds us today. "Narrow" might be a more apt descriptor for this type of AI as it is anything but weak: it enables some very robust applications, such as Apple's Siri, Amazon's Alexa, IBM watsonx™, and self-driving vehicles.

Strong AI is made up of artificial general intelligence (AGI) and artificial super intelligence (ASI). AGI, or general AI, is a theoretical form of AI where a machine would have an intelligence equal to humans; it would be self-aware with a consciousness that would have the ability to solve problems, learn, and plan for the future. ASI—also known as superintelligence—would surpass the intelligence and ability of the human brain. While strong AI is still entirely theoretical with no practical examples in use today, that doesn't mean AI researchers aren't also exploring its development. In the meantime, the best examples of ASI might be from science fiction, such as HAL, the superhuman and rogue computer assistant in 2001: A Space Odyssey.

Machine learning and deep learning are sub-disciplines of AI, and

deep learning is a sub-discipline of machine learning.

Both machine learning and deep learning algorithms use neural networks to 'learn' from huge amounts of data. These neural networks are programmatic structures modeled after the decision-making processes of the human brain. They consist of layers of interconnected nodes that extract features from the data and make predictions about what the data represents.

Machine learning and deep learning differ in the types of neural networks they use, and the amount of human intervention involved. Classic machine learning algorithms use neural networks with an input layer, one or two 'hidden' layers, and an output layer. Typically, these algorithms are limited to supervised learning: the data needs to be structured or labeled by human experts to enable the algorithm to extract features from the data.

Deep learning algorithms use deep neural networks—networks composed of an input layer, three or more (but usually hundreds) of hidden layers, and an output layout. These multiple layers enable unsupervised learning: they automate extraction of features from large, unlabeled and unstructured data sets. Because it doesn't require human intervention, deep learning essentially enables machine learning at scale.

Generative AI refers to deep-learning models that can take raw data— say, all of Wikipedia or the collected works of Rembrandt—and "learn" to generate statistically probable outputs when prompted. At a high level, generative models encode a simplified representation of their training data and draw from it to create a new work that's similar, but not identical, to the original data.

Generative models have been used for years in statistics to analyze numerical data. The rise of deep learning, however, made it possible to extend them to images, speech, and other complex data types. Among the first class of AI models to achieve this cross-over feat were variational autoencoders, or VAEs, introduced in 2013. VAEs were the first deep-learning models to be widely used for generating realistic images and speech.

"VAEs opened the floodgates to deep generative modeling by making models easier to scale," said Akash Srivastava, an expert on generative AI at the MIT-IBM Watson AI Lab. "Much of what we think of today as generative AI started here."

Early examples of models, including GPT-3, BERT, or DALL-E 2, have shown what's possible. In the future, models will be trained on a broad set of unlabeled data that can be used for different tasks, with minimal fine-tuning. Systems that execute specific tasks in a single domain are giving way to broad AI systems that learn more generally and work across domains and problems. Foundation models, trained on large, unlabeled datasets and fine-tuned for an array of applications, are driving this shift.

As to the future of AI, when it comes to generative AI, it is predicted that foundation models will dramatically accelerate AI adoption in enterprise. Reducing labeling requirements will make it much easier for businesses to dive in, and the highly accurate, efficient AI-driven automation they enable will mean that far more companies will be able to deploy AI in a wider range of mission-critical situations. For IBM, the hope is that the computing power of foundation models can eventually be brought to every enterprise in a frictionless hybrid-cloud environment.

There are numerous, real-world applications for AI systems today. Below are some of the most common use cases:

Speech recognition
Also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, speech recognition uses NLP to process human speech into a written format. Many mobile devices incorporate speech recognition into their systems to conduct voice search—Siri, for example—or provide more accessibility around texting in English or many widely-used languages.

Customer service
Online virtual agents and chatbots are replacing human agents along the customer journey. They answer frequently asked questions (FAQ) around topics, like shipping, or provide personalized advice, cross-selling products or suggesting sizes for users, changing the way we think about customer engagement across websites and social media platforms. Examples include messaging bots on e-commerce sites with virtual agents , messaging apps, such as Slack and Facebook Messenger, and tasks usually done by virtual assistants and voice assistants.

Computer vision
This AI technology enables computers and systems to derive meaningful information from digital images, videos and other visual inputs, and based on those inputs, it can take action. This ability to provide recommendations distinguishes it from image recognition tasks. Powered by convolutional neural networks, computer vision has applications within photo tagging in social media, radiology imaging in healthcare, and self-driving cars within the automotive industry.

Supply chain
Adaptive robotics act on Internet of Things (IoT) device information, and structured and unstructured data to make autonomous decisions. NLP tools can understand human speech and react to what they are being told. Predictive analytics are applied to demand responsiveness, inventory and network optimization, preventative maintenance and digital manufacturing. Search and pattern recognition algorithms—which are no longer just predictive, but hierarchical—analyze real-time data, helping supply chains to react to machine-generated, augmented intelligence, while providing instant visibility and transparency.

Weather forecasting
The weather models broadcasters rely on to make accurate forecasts consist of complex algorithms run on supercomputers. Machine-learning techniques enhance these models by making them more applicable and precise.

Anomaly detection
AI models can comb through large amounts of data and discover atypical data points within a dataset. These anomalies can raise awareness around faulty equipment, human error, or breaches in security.

The idea of "a machine that thinks" dates back to ancient Greece. But since the advent of electronic computing (and relative to some of the topics discussed in this article) important events and milestones in the evolution of artificial intelligence include the following:

1950: Alan Turing publishes Computing Machinery and Intelligence (link resides outside ibm.com). In this paper, Turing—famous for breaking the German ENIGMA code during WWII and often referred to as the "father of computer science"— asks the following question: "Can machines think?" From there, he offers a test, now famously known as the "Turing Test," where a human interrogator would try to distinguish between a computer and human text response. While this test has undergone much scrutiny since it was published, it remains an important part of the history of AI, as well as an ongoing concept within philosophy as it utilizes ideas around linguistics.
1956: John McCarthy coins the term "artificial intelligence" at the first-ever AI conference at Dartmouth College. (McCarthy would go on to invent the Lisp language.) Later that year, Allen Newell, J.C. Shaw, and Herbert Simon create the Logic Theorist, the first-ever running AI software program.
1967: Frank Rosenblatt builds the Mark 1 Perceptron, the first computer based on a neural network that "learned" though trial and error. Just a year later, Marvin Minsky and Seymour Papert publish a book titled Perceptrons, which becomes both the landmark work on neural networks and, at least for a while, an argument against future neural network research projects.
1980s: Neural networks which use a backpropagation algorithm to train itself become widely used in AI applications.
1995: Stuart Russell and Peter Norvig publish Artificial Intelligence: A Modern Approach (link resides outside ibm.com), which becomes one of the leading textbooks in the study of AI. In it, they delve into four potential goals or definitions of AI, which differentiates computer systems on the basis of rationality and thinking vs. acting.
1997: IBM's Deep Blue beats then world chess champion Garry Kasparov, in a chess match (and rematch).
2004: John McCarthy writes a paper, What Is Artificial Intelligence? (link resides outside ibm.com), and proposes an often-cited definition of AI.

2011: IBM Watson beats champions Ken Jennings and Brad Rutter at Jeopardy!
2015: Baidu's Minwa supercomputer uses a special kind of deep neural network called a convolutional neural network to identify and categorize images with a higher rate of accuracy than the average human.
2016: DeepMind's AlphaGo program, powered by a deep neural network, beats Lee Sodol, the world champion Go player, in a five-game match. The victory is significant given the huge number of possible moves as the game progresses (over 14.5 trillion after just four moves!). Later, Google purchased DeepMind for a reported USD 400 million.
2023: A rise in large language models, or LLMs, such as ChatGPT, create an
enormous change in performance of AI and its potential to drive enterprise value. With these new generative AI practices, deep-learning models can be pre-trained on vast amounts of raw, unlabeled data.