VASEM

VIRGINIA ACADEMY *of* SCIENCE, ENGINEERING, AND MEDICINE

An Introduction to Artificial Intelligence

Prepared by the Virginia Academy of Science, Engineering, and Medicine

Virginia Academy of Science, Engineering, and Medicine

The Virginia Academy of Science, Engineering, and Medicine is a nonprofit organization consisting of members of the National Academies of Science, Engineering, and Medicine who reside or work in Virginia as well as Virginians who are leaders in these fields. Through its nonpartisan network of experts, the Virginia Academy provides rigorous analytical, technical, and scientific support to inform policy on issues critical to the Commonwealth.

The Virginia Academy also promotes research, fosters interchange among individuals and organizations, and recognizes and honors Virginians who have made major contributions to science, engineering, and medicine.

Publishing and Writing: Charles Feigenoff
Editing: Rosalind Hingeley
Design: Roseberries
Cover Illustration: Alamy Stock Photos

Contents

WHITE PAPER

An Introduction to

Artificial Intelligence

Prepared by the Virginia Academy of Science, Engineering, and Medicine

VASEM

VIRGINIA ACADEMY of SCIENCE, ENGINEERING, AND MEDICINE

LOGIN / ALAMY

Understanding Artificial Intelligence

Artificial intelligence (AI) has reached an inflection point. Thanks to rapid advances over the last decade in computing hardware and software, artificial intelligence is in the process of transforming virtually every field of human endeavor, from medical research and healthcare to manufacturing and marketing. Automobile manufacturers are using AI-driven robots to assemble cars faster and with fewer defects than human workers can do. AI tools are helping physicians diagnose cancer quickly and accurately by analyzing medical images and patient records. Financial institutions use AI algorithms to detect fraud by scanning millions of transactions in real time and spotting unusual patterns that would otherwise be difficult to detect. Software developers are using AI to generate code and to streamline the process of updating legacy applications. Even more significant changes are in the offing.

Although AI is superior to human beings in harnessing massive amounts of data to develop summaries, generate responses, and make predictions, it has its limitations. AI still falls short, for instance, in its ability to adapt to new and unforeseen situations by transferring knowledge and skills learned in one domain to another, but it is moving fast in that direction. All the changes we have witnessed thus far are just a portent of AI's ability to transform society.

With the hope of fostering more informed discussion, the Virginia Academy of Science, Engineering, and Medicine has developed this introduction to promote general understanding of artificial intelligence. We include an account of important stages in the development of the field including key concepts and commonly used terms, the process by which generative AI chatbots like ChatGPT are trained, and the persistent challenge of eliminating hallucinations or erroneous responses. Finally, we provide an overview of some of the more prominent ethical issues that AI raises as well as offer readers an idea of what the future holds.

This paper has been reviewed by some of Virginia's leading AI experts:

Scott Acton, PhD, AI Advisor to the Provost, University of Virginia

Jeffrey Colombe, PhD, Principal Scientist: Neuroscience, AI/ML, Data Science, MITRE

Anuj Karpatne, PhD, Associate Professor and College of Engineering Faculty Fellow, Virginia Tech

Amarda Shehu, PhD, Vice President and Chief AI Officer, George Mason University

Aidong Zhang, PhD Thomas M. Linville Professor of Computer Science, University of Virginia

A LONG TIME COMING

Although AI seems to have burst into the public consciousness only recently, its roots go back 75 years. In 1950, British mathematician Alan Turing, who had played a critical role in cracking the German Enigma code machine during World War II, published his seminal paper, "Computing Machinery and Intelligence." In it, he proposed the idea that machines could eventually simulate human intelligence. He introduced the famous ***Turing Test***, using a machine's ability to conduct conversations indistinguishable from those of a human being as an indication of intelligent behavior. By that measure, AI has made great strides. At the end of March 2025, researchers at the University of California San Diego published a preprint study that found that conversations conducted with OpenAI's GPT-4.5 were mistaken for those with human beings more than 70% of the time.[1] To what extent the Turing Test is a useful or genuine measure of artificial intelligence is a matter of some debate.

CIA / ALAMY

The Enigma Machine

Just a few years later at the Dartmouth Conference in 1956, John McCarthy coined the term "artificial intelligence," which launched AI as a field of research. He defined AI as "the science and engineering of making intelligent machines." Claude Shannon, one of the conference's key participants, reportedly objected to McCarthy's coinage, in large part because of his belief that the term might create unrealistic expectations. Recently reignited debates about what the field can and will achieve reflect this early disagreement between McCarthy and Shannon.

Subsequent researchers found that realizing the goals of AI to be unexpectedly difficult. Given the Turing Test's emphasis on conversation, researchers in the 1960s and 1970s focused on ***natural language processing***, enabling computers to interpret and generate text that resembles the way human beings naturally speak or write. Notable examples include ELIZA, a program that simulated a psychotherapist, and SHRDLU, one that used natural language prompts to manipulate objects in a virtual world.

[1] Cameron Jones and Benjamin Bergen, "Large Language Models Pass the Turing Test," ArXiv 2503.23674v1 (in press), https://arxiv.org/pdf/2503.23674.

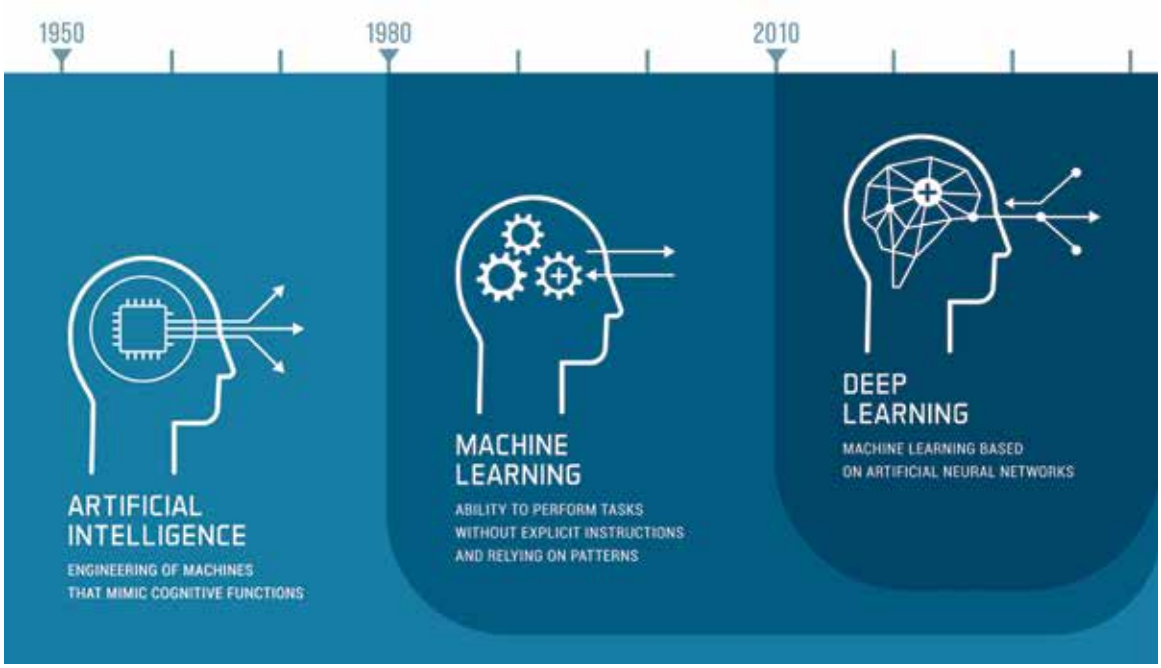


Figure 1: Advances in Artificial Intelligence

These programs, however, worked only in rigidly defined situations. ELIZA, for instance, employed ***rule-based pattern matching*** and templated responses. A user might say, "I feel sad today." ELIZA would always answer, "Why do you feel sad today?" Although SHRDLU was more flexible, it existed in a drastically simplified digital world and rigidly applied the rules of grammar and logic to analyze and respond to prompts.

The limits of both programs underscored the drawbacks of the rule-based approach to natural language processing as well as the need for vastly more powerful computers able to simulate real-world complexity. Furthermore, these programs were not capable of ***machine learning***, the ability of a program to improve from experience. Artificial intelligence had reached an impasse. (See Figure 1 for steps in the development of AI.)

THE EMERGENCE OF EXPERT SYSTEMS

The field regained momentum in the 1980s with the rise of expert systems, programs that combine a knowledge base and a set of rules to analyze information and solve problems in a carefully defined field. A breakthrough program of the period was XCON, developed by Carnegie Mellon University for the Digital Equipment Corporation (DEC). XCON could automatically configure DEC's VAX computer systems to customer specifications, a complex and time-consuming process. When a customer ordered a DEC VAX computer, XCON took customer requirements and

generated the correct configuration, including hundreds of parts, connections, and settings. XCON dramatically reduced errors, delays, and cost.

Although expert systems of the period like XCON could solve complex problems quickly and consistently, they did not learn from experience. They were also poor at handling new or unusual situations and were hard to update as knowledge evolved. XCON was based on more than 2,000 rules that distilled the expertise of human configurators, but manually encoding new rules was time-consuming, and XCON became more brittle as the number of rules grew. These early expert systems were a revelation, but they were rigid. Research stalled again.

MACHINE LEARNING TAKES OFF

Progress in AI revived in the early 1990s, thanks in part to the massive growth of digital data generated by the emerging Internet. This surge of easily available data created the perfect environment for machine learning, which requires massive datasets for training. At the same time, ***Moore's Law***, which predicts the doubling of transistors in a computer chip roughly every two years, facilitated the shift from rules-based approaches to open-ended ***algorithms***, instruction sets in mathematical format for solving a problem or performing a computation. These algorithms, which rely on probability theory and statistics, make active machine learning possible. Researchers began to move from hard-coding rules to giving machines the power to generate responses based on their analysis of large quantities of data.

AI began to be deployed for such real-world applications as speech recognition, optical character recognition, credit scoring, and medical diagnosis support. But the power of AI was most dramatically demonstrated in a decidedly noncommercial setting. In 1997, IBM's ***Deep Blue*** defeated world chess champion Garry Kasparov. Deep Blue used brute-force computing rather than learning to master the game. When it was its turn, Deep Blue ran through 200 million possible sequences of chess moves per second and chose the one that would lead to the best outcome. Its victory highlighted the potential of specialized AI systems to surpass human abilities at clearly defined tasks.

IEVGEN SKRYPKO / ALAMY

TRAINING TECHNIQUES FOR MACHINE LEARNING

Developers use several techniques to train machine learning models. In ***supervised learning*** (see Figure 2), programmers use labeled training data. For instance, they might train an autonomous driving program with labeled images of signs, lanes, and pedestrians. The AI model learns to identify unlabeled versions of these inputs and label them correctly. Since labeled training data are hard to obtain in many fields, an important specialization in supervised learning is training models to be accurate using a limited number of labeled samples.

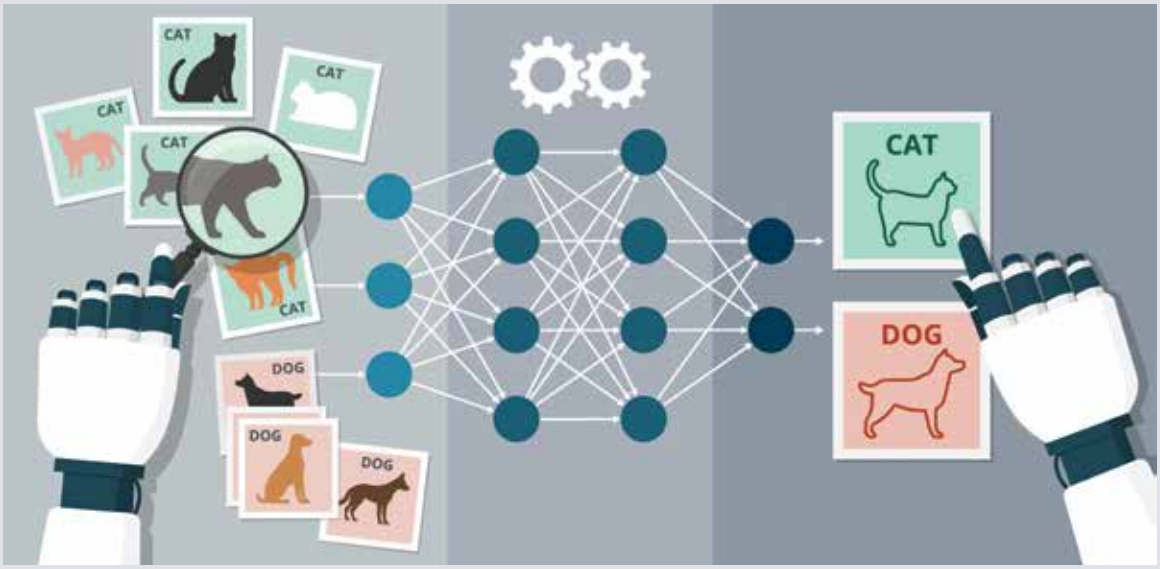


Figure 2: Advances in Artificial Intelligence

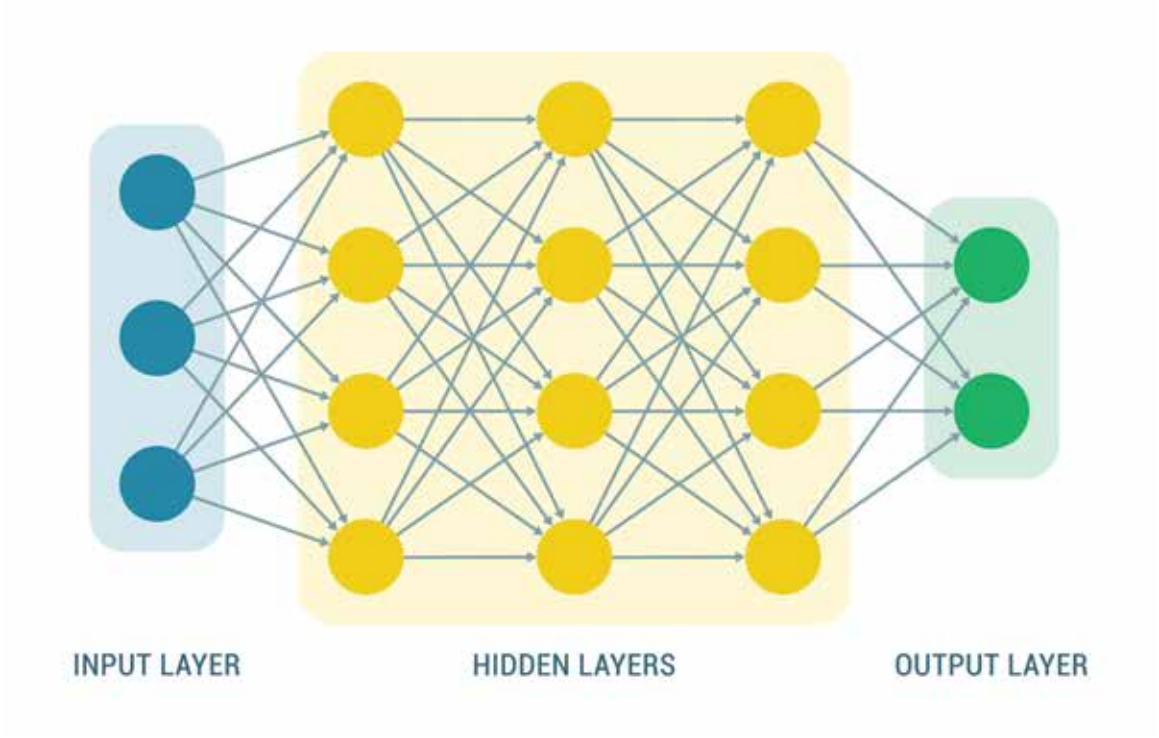
By contrast, ***unsupervised learning*** uses unlabeled data to train models. The purpose of unsupervised learning is to teach models to find hitherto unknown dependencies, regularities, or patterns among input values rather than generate a paired response to a single input. E-commerce platforms, for instance, are trained using unsupervised clustering methods to group customers with similar purchasing or browsing behaviors and use that information to generate product recommendations tailored to their preferences.

Another alternative for training models is ***reinforcement learning***, which uses human feedback about an AI model's performance to improve its performance. The feedback signal takes the form of a number similar to the "you're getting warmer" or "you're getting colder" advice in a treasure hunt. Reinforcement learning encourages desirable outputs and discourages undesirable ones.

THE ADVENT OF DEEP LEARNING

In the 1990s and 2000s, researchers returned to a decades-old line of AI research inspired by neuroscience and cognitive science to advance the field of machine learning. They pursued the development of ***artificial neural networks***, which are loosely modeled after the way the human brain processes information. Neural networks that were typical of this era of research consist of nodes (the equivalent of neurons) organized into three main layers: an input layer that reads raw data, hidden layers that process the data using weighted connections and nonlinear functions, and an output layer that makes predictions based on this data. Each connection has a weight—and the network adjusts those weights during training to reduce error in its predictions. Figure 3 shows how a basic deep learning neural network is organized, and Figure 4 shows a network after weights are adjusted. When trained on massive amounts of unlabeled data, neural networks learn to identify intrinsic patterns in the data and, most significantly, can also generalize when presented with new data.

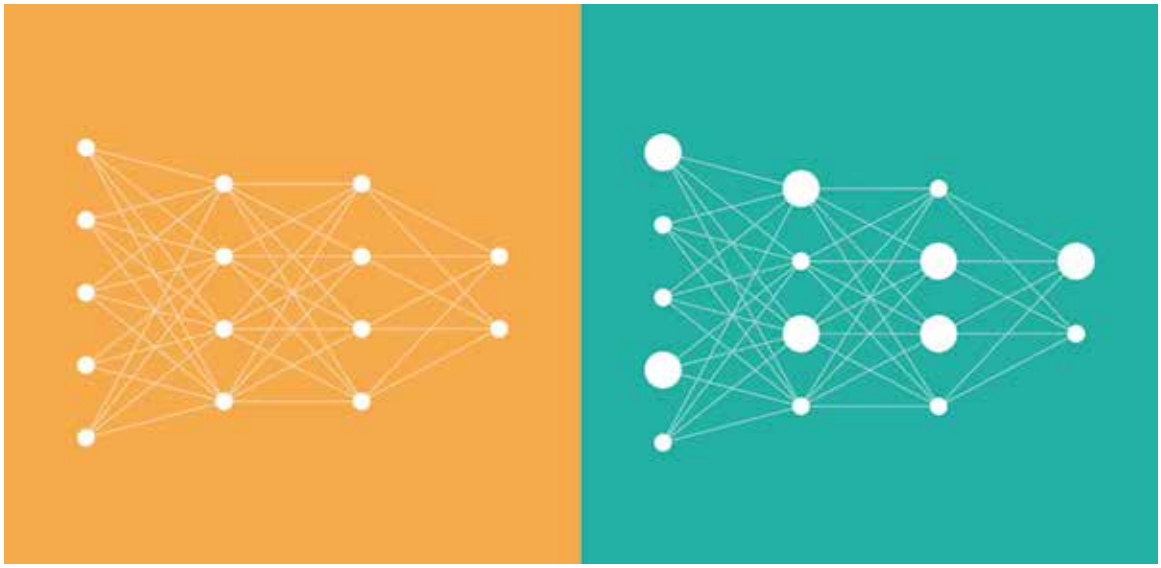
Deep learning networks have a series of hidden layers, ranging from a handful to several thousand in state-of-the-art models, that build on each other. For instance, a model might generate a response by moving through layers that embody increasingly complex and abstract concepts. When asked to recognize an image, the network might move through a progressive series of hidden layers to first identify edges, then shapes, and then objects. Deep learning models can automatically extract arbitrarily complex features from data, earning them the name "universal feature approximators."



ELENABS/ ALAMY

Figure 3: The Organization of a Basic Deep Learning Neural Network

In the 2012 ImageNet competition, a deep learning neural network developed by researchers at the University of Toronto and trained using supervised learning techniques dramatically outperformed traditional methods of image classification. Deep learning networks have also been applied to problems without using supervised learning. For example, in 2016 Google Deep Mind's ***AlphaGo*** used deep learning networks trained using reinforcement learning to beat Lee Sedol, a world master of the complex board game Go. This is an achievement once thought impossible for AI.



ELENABS/ ALAMY

Figure 4: Untrained and Trained Neural Network Showing Weighted Nodes

GENERATIVE AI

Deep learning forms the foundation of powerful ***large language models (LLMs)*** like OpenAI's ChatGPT, Google Deep Mind's Gemini, and Anthropic's Claude. They can interact with human users using natural language. The foundational architecture for most LLMs is the transformer, which is designed to work with sequential systems like text, music, and code. Transformers feature a self-attention mechanism that allows the model, for instance, as it looks for the next word in a sequence of text, to weigh the importance of different words in the sequence relative to each other, helping it understand the relationships between them. LLMs use this learned ability to generate long multiword answers in response to specific typed queries from human users. These queries are called ***prompts***.

LLMs are examples of ***generative AI***, which goes beyond making predictions—for instance, about whether a patient with a particular mammogram has breast cancer—to generating new content. In responding to a prompt, it uses the structure and distribution embedded in the data itself to produce an answer.

Progress in generative AI has dramatically accelerated in the past few years, thanks in part to the development of ***self-supervised learning*** techniques. Instead of using costly labeled datasets to improve predictive accuracy on a single task, programmers create pretext tasks for models to solve using unlabeled data. In order to solve these tasks, the model must identify the inherent structure and patterns within the data.

Generative AI models such as LLMs are trained using a "pretrain-then-fine-tune paradigm." As part of the pretraining phase, LLMs are fed huge quantities of unlabeled text data. Self-supervised learning is used to train them to learn grammar, reasoning patterns, and general knowledge. During fine-tuning, developers add a layer of human-labeled examples, for instance, examples of pairs of questions and answers to train LLMs on a variety of downstream tasks. This step (also referred to as instruction fine-tuning) trains models to be effective at answering questions and attempting to satisfy user requests. Developers also use reinforcement learning to enable models to produce better responses over time.

Major Large Language Models

Model	*Developer*
GPT	OpenAI
Gemini	Google DeepMind
Claude	Anthropic
LLaMA	Meta AI
Mistral	Mistral AI
Command	Cohere
Grok	xAI
DeepSeek	DeepSeek

Other recent advances in training include ***few-shot learning***, ***zero-shot learning***, and ***in-context learning***. In few-shot learning, a machine learning model needs to be exposed to only a few examples to learn the patterns that determine its output. For example, to train a program to translate from English to French, the developer might provide a few examples of paired English and French phrases. The program then uses its previous language knowledge and these examples to find the French equivalent for any word in English. By contrast, a zero-shot learning algorithm could perform this task without previous examples, relying on a clear instruction and its internal understanding of the elements of translation derived from its broad pretraining.

In-context learning is a rather new phenomenon, a somewhat surprising revelation of inherent capabilities exhibited by large language models. LLMs learn to perform a task by using a few examples provided directly within the input prompt instead of undergoing explicit fine-tuning on a separate dataset. Developers can train a model that uses the context or examples in a prompt to infer what the user wants without having to modify its internal parameters through fine-tuning. The resulting speed and flexibility make large language models increasingly useful as conversational agents.

JIM WEST/ ALAMY

An AI-Powered Robotic Bin-Picker

REPRESENTATIVE APPLICATIONS OF AI

Given the power that AI has achieved, it is not surprising that it is being embedded in virtually every human endeavor. AI is transforming a wide range of sectors by automating tasks, improving decision-making, and enabling new capabilities. Here are several representative applications of AI:

Manufacturing. AI is being used to predict demand, optimize logistics, and anticipate equipment failure. ***Computer vision***, a sub-field of AI that enables machines to interpret and understand images, videos, and real-time camera feeds, enables robots to inspect products for defects. Computer vision also allows robots to optimize motor control, making it possible for them to assemble delicate parts.

Astronomy. Modern space- and Earth-based observatories generate terabytes of data every day. AI is used to filter out irrelevant data to help detect exoplanets, pulsars, and supernovae. It also enables spacecraft and rovers to navigate autonomously and to assist the development of simulations that model galaxy formation, black holes, and dark matter distribution.

Healthcare. AI is accelerating drug development by simulating how different molecules interact with targets in the body and by proposing new drug candidates. It is also being used to model disease outbreaks, analyze population health data, and assist in resource planning. At the clinic, AI-powered systems, trained on thousands of medical papers and clinical records, analyze X-rays, MRIs, CT scans, and pathology slides to detect abnormalities quickly and accurately. They also provide clinical decision support to doctors by recommending treatment options and flagging potential complications.

Finance. Hedge funds have made extensive use of deep learning models to monitor markets, economic indicators, and even social media sentiment. They are also used to identify emerging arbitrage opportunities and execute trades in milliseconds. Other financial firms employ AI to analyze market trends, forecast returns, and optimize investment portfolios based on investor risk profiles.

National Security and Defense. AI is transforming national security by enhancing threat detection, streamlining intelligence analysis, and enabling precise military operations by incorporating AI-enabled autonomy into reconnaissance, logistics, and battlefield awareness systems. These technological advances are reshaping military doctrine and security strategies, creating new capabilities while generating complex ethical questions.

THE BLACK BOX: HALLUCINATIONS AND TRUST

One of the major challenges of using artificial neural networks is that the neurons and connection weights that make up the network do not have preassigned values, and no labels or interpretations are assigned to specific internal elements during training. This has led many researchers and practitioners in the field to regard artificial neural networks as uninterpretable black boxes whose inner workings are mysterious. The trustworthiness of models whose mechanisms of operation cannot be inspected, interpreted, or reasoned with cannot be established.

Their opacity poses a dilemma for their developers. How can developers guarantee the accuracy of responses if they do not understand how responses are generated? A model simply applies algorithms to the data that has been used to train it. It is nearly impossible to know what combination of specific neurons or layers generates a specific false response and whether the fault lies with the underlying algorithms, deficiencies in the training data, the construction of the prompt, or some combination of all three.

Large Learning Model Hallucinations

Fake Legal Citations Two New York lawyers were sanctioned for submitting a brief containing fictitious citations drafted by ChatGPT. The model cited a case that did not exist, *Varghese v. China Southern Airlines Co., Ltd.*, 925 F.3d 1339 (11th Cir. 2019), and provided extracts from the judgement. When questioned, ChatGPT doubled down, explaining that the case "does indeed exist and can be found on legal research databases such as Westlaw and LexisNexis."[2]

False Accusations When asked about the role that Brian Hood, mayor of the Hepburn Shire Council in Victoria, Australia, had in a financial scandal, ChatGPT responded that he was imprisoned for bribery while working for a subsidiary of the country's national bank. In fact, Hood was a whistleblower and was never charged with a crime.[3]

Incorrect Product Information A chatbot used by Air Canada assured a customer he could book a full-fare flight for his grandmother's funeral and then apply for a bereavement fare later. When he did, the airline refused to issue a discount telling him that the request had to be submitted before the flight. The airline lost in court, paying damages to the customer and court costs.[4]

Furthermore, while the ability of large language models to create novel content is a source of their strength, on occasion they generate false responses—for instance, designating the wrong location for an event—or make up responses. These are called ***hallucinations***. ChatGPT, for instance, has invented legal precedents, created false, compromising narratives about individuals, invented historical events and scientific facts, and made misleading claims about products and services. Large language models cannot distinguish between what is true or false. They are simply trained to use statistical processes to generate plausible responses.

To address these issues, developers are employing a number of approaches. The companies behind the major large learning models are connecting them to real-time databases or search engines to improve their reliability. This approach is known as ***retrieval augmented generation***. Researchers at

[2] "Artificially Unintelligent: Attorneys Sanctioned for Misuse of ChatGPT - Insights - Proskauer Rose LLP," Proskauer, n.d., https://www.proskauer.com/blog/artificially-unintelligent-attorneys-sanctioned-for-misuse-of-chatgpt.

[3] "LLM Hallucinations: Complete Guide to AI Errors," SuperAnnotate (blog), August 8, 2024, https://www.superannotate.com/blog/ai-hallucinations#:~:text=In%20the%20same%20month%2C%20ChatGPT,not%20always%20about%20individual's%20reputation.

[4] Maria Yagoda, "Airline Held Liable for Its Chatbot Giving Passenger Bad Advice - What This Means for Travellers," February 23, 2024, https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know.

universities and in industry are also trying to penetrate the black box. Among other initiatives, they are working to understand how specific neurons and layers in a model contribute to its behavior, attempting to track the parts of a model's training data that influenced a specific output, and seeking to understand what information is stored and processed in specific neurons or layers.

In 2024 and 2025, developers began to introduce specialized reasoning systems. For instance, in April 2025, OpenAI introduced o3 and o4-mini, which are available to subscribers of ChatGPT. They are called reasoning systems because they generate a ***chain of thought*** before answering. A chain of thought is a technique that prompts large language models to explain their reasoning process step-by-step. Reasoning systems take longer than other models but provide better results for science and mathematics applications.

FENG YU/ ALAMY

ETHICAL ISSUES

The transformational power of AI has also generated widespread concern. AI is a massively disruptive technology and threatens to change the nature of work for employees at all levels of the economy: designers and journalists as well as truck drivers, call center agents, and factory workers. As the practical applications of generative AI spread, the range of affected employees—knowledge workers as well as manual laborers—will grow.

AI also can be manipulated to foster misinformation and disinformation. AI-powered deepfake technology can be used to generate convincing videos that can sway elections or encourage violence. It is also being used for scams. In 2024, forensics and identification verification company Regula found that half of all its business customers had experienced fraud involving audio and video deepfakes, costing them an average of nearly $450,000 each.[5]

[5] "How to Fight AI-Boosted Spear Phishing Fraud," Independent Banker, March 1, 2025, https://www.independentbanker.org/article/2025/03/01/how-to-fight-ai-boosted-spear-phishing-fraud#:~:text=During%20a%20video%20call%2C%20the,to%20trick%20the%20Arup%20employee.

A $25 Million Scam

In 2024, an employee of the multinational design and engineering company Arup was duped into attending a video call with people she believed to be the chief financial officer and the employee's colleagues, but who were actually deepfakes. At the supposed chief financial officer's request, the employee sent $25 million to bank accounts controlled by criminals.[6] Arup's chief information officer has used the occasion to publicize similar incidents, which are more common than generally appreciated.

In addition, AI has the potential to perpetuate existing patterns of bias and discrimination in such tasks as medical diagnoses, surveillance, and hiring. For instance, an AI recruiting system used by Amazon was found to favor male candidates over females because it had been trained on historical data in which men dominated tech roles.[7] Bias comes in multiple forms. ***Value bias*** is a conscious or unconscious preference for one thing over another (e.g., Dogs are clearly better than cats). ***Statistical bias*** is a factual preponderance of existing data that may run counter to one's cultural values or one's desired vision of the future (e.g., Most nurses are female). ***Sampling bias*** is an uneven access to data that can be misleading (e.g., Ninety percent of people surveyed say they don't mind volunteering to take surveys). Each kind of bias can introduce its own forms of failure in AI models. The aura of algorithmic objectivity surrounding high-tech systems tends to obscure their weaknesses, but a growing subdomain of AI research focuses on unraveling and mitigating sources of bias in models and the data used to train them.

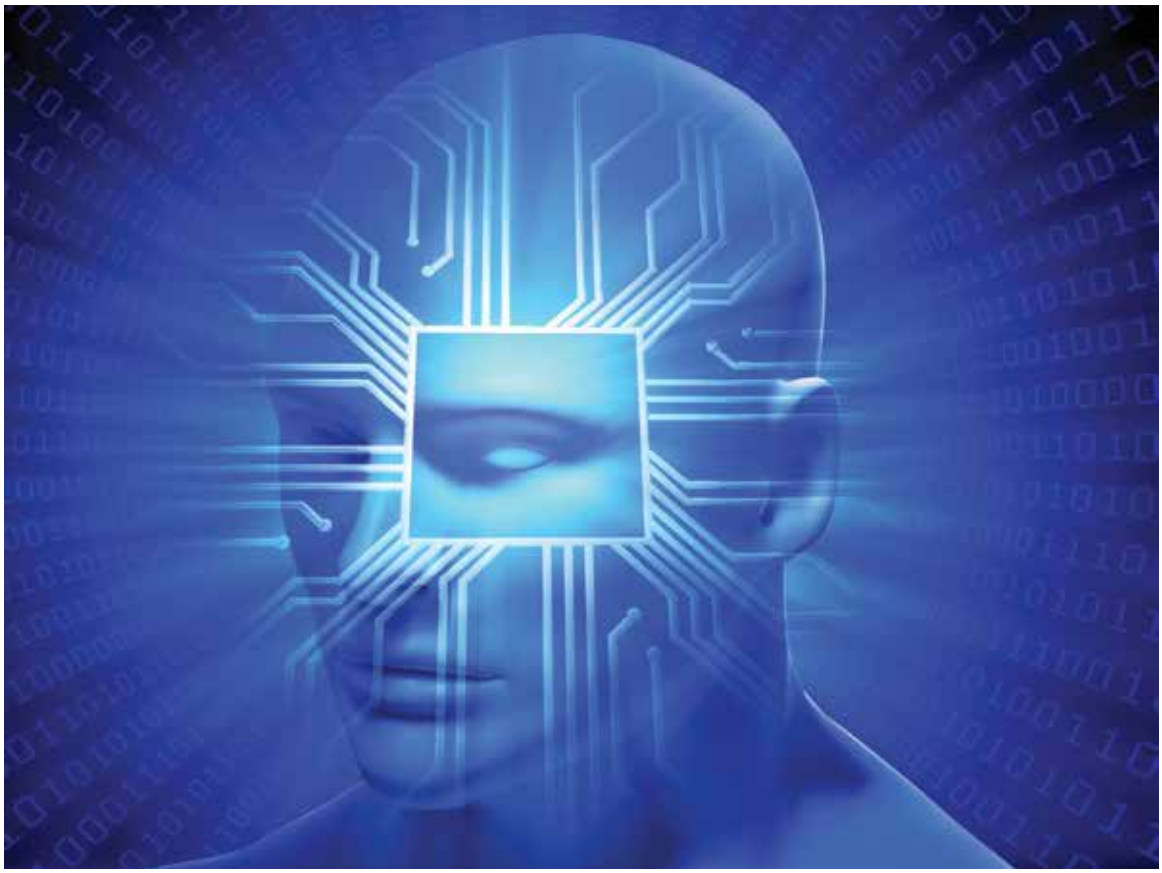
Finally, disputes about the indiscriminate use of intellectual property to train AI have become widespread. *The New York Times* sued OpenAI and Microsoft in December 2023, arguing that their large language models unlawfully incorporated the newspaper's copyrighted articles into their training datasets. The lawsuit claims that the companies used millions of the newspaper's articles without permission to help train chatbots.[8]

---

[6] Kathleen Magramo, "British Engineering Giant Arup Revealed as $25 Million Deepfake Scam Victim," cnn.com, May 17, 2024, https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk/index.html.

[7] Jeffrey Dastom, "Insight - Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, October 10, 2018, https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/.

[8] Negar Bondari, "AI, Copyright, and the Law: The Ongoing Battle Over Intellectual Property Rights – IP & Technology Law Society," February 4, 2025, https://sites.usc.edu/iptls/2025/02/04/ai-copyright-and-the-law-the-ongoing-battle-over-intellectual-property-rights/.



ZOONAR GMBH/ ALAMY

ARTIFICIAL INTELLIGENCE VS. HUMAN INTELLIGENCE

Despite the tremendous advances during the past decade, artificial intelligence is far from being interchangeable with human intelligence. Although it can reason about itself, it has no consciousness or understanding. In an essay about AI and education, *New Yorker* author Graham D. Burnett assigned his students an exercise that led them to query ChatGPT on the concept of comprehension. The model replied to one student's query:

> "Even though I can generate text that sounds like understanding, my process doesn't involve the internal experience of meaning. Humans comprehend because they synthesize information into a unified, lived experience—they feel, they interpret, they reflect. I don't. I process, predict, and structure, but there is no subjective experience underlying my words."[9]

[9] D. Graham Burnett, "Will the Humanities Survive Artificial Intelligence?" *The New Yorker*, April 26, 2025, https://www.newyorker.com/culture/the-weekend-essay/will-the-humanities-survive-artificial-intelligence.

In other words, at least according to ChatGPT, large language models can provide detailed answers to questions or verbally depict complex events at great length, but they may do so without having access to a deeper understanding of what they are talking about. The ability of large language models to use language with a degree of competence that in some ways approaches or surpasses human ability does not imply parity with humans in all aspects of cognition and experience.

THE NEXT STEPS

AI remains a work in progress—and there are scores of issues yet to be addressed. For instance, many tests enable users to evaluate the power and accuracy of different AI systems, but none tell which model is best for a specific use case. Should you use OpenAI's ChatGPT for coding, or is Google's Gemini a better option? Existing benchmarks measure performance in very specific contexts that are often found to be divorced from real-world applications. The commonly used massive multitask understanding (MMLU) test consists of tens of thousands of multiple-choice academic questions across a range of subjects. The assumption is that a model with a higher score is smarter than one that does less well, but there is no guarantee that this translates into excellence at a specific task.[10]

In fact, some accuse model providers of manipulating benchmarks to artificially inflate their scores, potentially skewing comparisons and hindering real-world performance. The recent controversy that erupted around Meta's Llama 4 Maverick model is an example. Critics accused the company of using customized training and optimization designed specifically to produce high scores on the LMArena, a popular AI benchmark.[11] This is the AI version of teaching to the test.

There are also broader issues that must be tackled if AI is to be more generally applicable. AI has been primarily fueled by the requirements of solving problems in two application domains—computer vision and natural language modeling—that benefit from the deluge of Internet-scale data. AI is not equally adept at modeling structured data used in scientific fields, such as spatial and temporal data in climate science and fluid dynamics or graph-based data in chemistry. Traditional deep learning approaches to scientific domains often produce physically inconsistent results and struggle to generalize to new data.

As a result, there is a growing interest in developing ***knowledge-guided machine learning*** approaches that leverage decades—and in many cases centuries—of accumulated scientific knowledge and combine it with machine learning techniques. In these systems, the representation of domain knowledge, from physical and simulation models to symbolic and logical representations, is integrated into machine learning models.

[10] Logan Kugler, "How Do You Measure AI?" Communications of the ACM, March 20, 2025, https://doi.org/10.1145/3708972.
[11] Collinear AI's Blog, "Gaming the System: Goodhart's Law Exemplified in AI Leaderboard Controversy," May 15, 2025, https://blog.collinear.ai/p/gaming-the-system-goodharts-law-exemplified-in-ai-leaderboard-controversy?utm_campaign=post&utm_medium=web

Another limitation of AI is inherent in its reliance on past data for training. Its predictive abilities are "based on probabilistically sampling past associations or existing correlations from its training data, with an eye toward likely and probable outcomes."[12] While this approach does not prevent AI systems from generating novel content and synthesizing across disparate sources of available information, it does suggest constraints on their expertise based on the limitations inherent in working from the past.

THE ULTIMATE FRONTIER: ARTIFICIAL GENERAL INTELLIGENCE

Human cognitive abilities are limited by exposure to available information, and human cognition is plagued by cognitive biases that interfere with accuracy in making judgments. However, human cognition involves the use of imagination to manipulate mental models and to consider "what if" scenarios, unconscious processing that can sometimes result in apparently intuitive leaps, the identification of analogies across very disparate domains of expertise, and transfer of learning across domains.

Researchers hope to achieve similar abilities in AI systems. Artificial general intelligence (AGI) refers to the idea of creating a form of AI that is at least as capable as human intelligence in a wide variety of cognitive tasks. Although there is no single definition of AGI, the general consensus is that a model possessed of its power would be able to learn and reason in multiple domains, adapt to new situations, understand context, and perform a wide range of tasks with a high degree of autonomy. AGI systems would have the ability to learn from experience and adapt their behavior based on new information. And they would have a vast repository of knowledge about the world that would allow them to make common-sense decisions.

One step in this direction is foundation models. Foundation models better capture the broader capability of learning over massive datasets and are capable of solving a variety of tasks with minimal additional task customization or fine-tuning. DALL-E, SORA, and other foundation models are trained over not just one modality of data (such as text), but various modalities (text and images, or text and video) and can generate content across data modalities.

[12] Mari Sako and Teppo Felin, "Does AI Prediction Scale to Decision Making?" Communications of the ACM, March 21, 2025, https://doi.org/10.1145/3722138.

THREE VISIONS FOR AN AI FUTURE

Researchers disagree about how long it will take to develop true AGI. Right now, it exists as an ideal rather than a specific set of parameters. They also debate about the implications of its emergence. Forecasters fall roughly into three camps.

The first group is epitomized by two computer scientists at Princeton, Sayash Kapoor and Arvind Narayanan, who have released an influential paper titled "AI as Normal Technology."[13] They argue that AI is like other powerful technologies, and the same human, organizational, and institutional inertia that slowed innovations like electrification will give human beings the time to adjust to AI's transformative economic and societal impacts.

Commentators in the second group link AI to intelligence augmentation (IA), in other words, to providing human beings the information they need to make better decisions. Some of these observers believe that in the natural course of things, artificial intelligence will evolve slowly, and its limitations will keep it from moving beyond intelligence augmentation for decades to come.

The final group is exemplified by Daniel Kokotajlo, formerly an AI-safety researcher working at Open AI and executive director of the AI Futures Project. He warns that it will be difficult to guide the evolution of AI and to maintain control over it. The nonprofit recently published "AI 2027,"[14] a detailed worst-case scenario in which superintelligent AI systems either dominate or exterminate the human race by 2030.

In many respects, these views say as much about their proponents' vision of social organizations as they do about the inherent qualities of AI. Do we see AI as a break with the past or an extension of it? Where it actually falls on that continuum will determine our future.

[13] Arvind Narayanan & Sayash Kapoor, "AI as Normal Technology," Knight First Amendment Institute, April 15, 2025, https://knightcolumbia.org/content/ai-as-normal-technology.

[14] Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean, "AI 2027," AI Futures Project, n.d., https://ai-2027.com/.

2025 VASEM BOARD OF DIRECTORS

JAMES AYLOR, PhD
Past President
Dean Emeritus, School of Engineering and Applied Science and the Louis T. Rader Emeritus Professor of Electrical and Computer Engineering, University of Virginia.

ANTHONY BEASLEY, PhD
Director, National Radio Astronomy Observatory.

BARBARA BOYAN, PhD
Former Dean, College of Engineering; Executive Director, Institute for Engineering and Medicine; and the Alice T. and William H. Goodwin Chair in Biomedical Engineering, Virginia Commonwealth University.
Member, National Academy of Engineering.

ROBERT CAREY, MD
Dean Emeritus, School of Medicine and the David A. Harrison III Distinguished Professor of Medicine, University of Virginia.
Member, National Academy of Medicine.

PATRICIA DOVE, PhD
University Distinguished Professor and the C.P. Miles Professor of Science, Virginia Tech.
Member, National Academy of Sciences.

ANTONIO ELIAS, PhD
Former Executive Vice President and Chief Technical Officer, Orbital ATK.
Member, National Academy of Engineering.

ALFRED GRASSO
President
Former President and CEO, MITRE Corporation.

ANITA JONES, PhD
Professor of Computer Science Emerita, University of Virginia.
Former Director of Defense Research and Engineering, U.S. Department of Defense.
Member, National Academy of Engineering.

ROBERT KAHN, PhD
President & CEO, Corporation for National Research Initiatives.
Member, National Academy of Engineering, National Academy of Sciences.

ALEX KRIST, MD
Inaugural Sheldon M. Retchin Professor in Healthcare Innovation, Virginia Commonwealth University.

CHEN-CHING LIU, PhD
The American Electric Power Research Professor and Director Emeritus of the Power and Energy Center, Virginia Tech.
Member, National Academy of Engineering.

GENERAL LESTER L. LYLES (Ret.)
Former Commander, Air Force Materiel Command.
Member, National Academy of Engineering.

X. J. MENG, MD, PhD
Second Vice President
University Distinguished Professor of Molecular Virology, Virginia Tech.
Member, National Academy of Sciences.

ROBERT PHILLIPS, MD
Vice President of Research and Policy, American Board of Family Medicine.
Member, National Academy of Medicine.

DAVID ROOP
First Vice President
Former Director of Electric Transmission, Dominion Energy.
Member, National Academy of Engineering.

JENNIFER WEST, PhD
Secretary/Treasurer
Dean, School of Engineering and Applied Science and Saunders Family Professor of Engineering and Applied Science, University of Virginia.
Member, National Academy of Engineering, National Academy of Medicine.

A. THOMAS YOUNG
Former Executive Vice President, Lockheed Martin.
Former Mission Director of Goddard Space Flight Center, NASA.
Member, National Academy of Engineering.

VIRGINIA ACADEMY OF SCIENCE, ENGINEERING, AND MEDICINE
Gateway Plaza
800 East Canal Street, 11th Floor
Richmond, VA 23219
www.VASEM.org
infor@vasem.org

vasem.org