CometCo

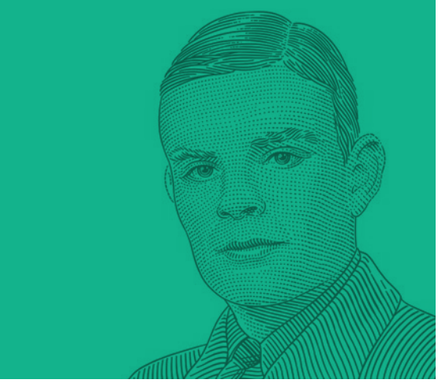
Hewlett Packard Enterprise

How AI and HPC Are Advancing Everything from Chip Design to Healthcare to Engineering

When Machines Think Big

"What we want is a machine that can learn from experience."

Alan Turing



Ever since ChatGPT burst onto the scene in 2022 as the most downloaded internet app ever, people can't seem to stop chattering about artificial intelligence (AI). Nonetheless, it has taken a long time for AI to mature to a point where it has the potential to revolutionize many fields of science and industry.

In the early part of the 20th century, a British mathematician, Alan Turing, theorized a futuristic machine "capable of computing any computable sequence." It would use tape for memory, and a head would move back and forth along the tape, reading and writing symbols. The same tape would also store a table of instructions, essentially a computer program that made everything work.

Turing had not not actually intended to build a computer — just to answer some fundamental mathematical questions. Nonetheless, his "Turing Machine" created the basis for all computers we use today.

By 1946, having helped the allies win World War II by cracking the Enigma code — which German armed forces used to send encrypted messages — Turing was already looking forward to one of the biggest advances in computer science. "What we want is a machine that can learn from experience, and that possibility of letting the machine alter its own instructions provides the mechanism for this." explained Turning. Unfortunately, the earliest computers lacked a key prerequisite for intelligence: you could tell them what to do, but they couldn't remember what they did. They lacked efficient memory.

The second hurdle to making machines think and communicate with humans was the extremely limited processing power of systems available at that time. Plus, the scant power they did have cost a king's ransom. It wasn't until after the second World War — as computers capable of executing commands and storing large quantities of data emerged — that the full potential of smart machines entered the range of possibility. The first computer with meaningful processing capabilities, UNIVAC (Universal Automatic Computer), went into service in 1951. Since the transistor had not been invented, its computer brain used thousands of vacuum tubes. Weighing more than seven tons, it could perform approximately 1900 operations per second. Compare that with the first supercomputer, the Cray 1 introduced in 1976, which could perform 1.5 million operations per second, and the central processing unit in the iPhone 16, which screams through data at speeds up to 35 trillion operations per second.

The first general purpose computer designed for business and research purposes, UNIVAC, went into service at the US Census Bureau in 1951.

Space porcupine? UNIVAC relied on tanks of liquid mercury to store data.

Also, UNIVAC had memory capable of keeping up with its processing power. Without the transistors and integrated circuits that make today's high-speed memory possible, UNIVAC used tanks of liquid mercury kept at 149 degrees F. Crystal transducers in each tank transmitted and received data as waves in the mercury.

UNIVAC achieved fame when it predicted the election victory of Republican Dwight D. Eisenhower in his presidential run against Democratic Illinois Gov. Adlai Stevenson. CBS news put it on live television on election night as Walter Cronkite anchored his first election night broadcast, as CBS reporter Charles Collingwood kept constant watch on UNIVAC. "This is the face of a UNIVAC," Collingwood told the CBS audience. "A UNIVAC is a fabulous electronic machine, which we have borrowed to help us predict this election from the basis of early returns as they come in." That rudimentary ability to predict an outcome from data represented one of the first baby steps toward AI.

8.30 P.M.

IT'S AWFULLY EARLY, BUT I'LL GO OUT ON A LIMB.

UNIVAC PREDICTS--with 3,398,745 votes in--

	STEVENSON	EISENHOWER
STATES	5	43
ELECTORAL	93	438
POPULAR	18,986,436	32,915,049

THE CHANCES ARE NOW 00 to 1 IN FAVOR OF THE ELECTION OF EISENHOWER.

A printout of the UNIVAC's prediction that Dwight Eisenhower would win the 1952 US presidential election.

Thinking ahead

Essentially, Turing foresaw what today we call machine learning and artificial intelligence. For a machine, the 'experience' described by Turing comes in the form of data. "When a machine learns from data and becomes smart enough to make an informed choice, we say the machine is artificially intelligent," explains Dr. Eng. Lim Goh Senior Vice President for Value from Data and AI, at Hewlett Packard Enterprise. "And unlike in basic computing, it does not need to be programmed by a human to make choices, but instead learns to make choices from data."

With the advent of transistors, semiconductors, large-scale integrated circuits, and larger data storage capacity, machines began to become smarter. Famously, in 1997 'Deep Blue,' a specialized computer built by IBM (playing on IBM's nickname of Big Blue), beat the reigning world chess champion, Garry Kasparov, in a six-game match. But there was debate as to whether Deep Blue was truly intelligent or was it simply channeling human intelligence through a program coded by the programmers. Turing had foreseen this question, and in 1957 proposed an artificial intelligence test that is now known as the Turing Test. It requires three participants: a computer, a human interrogator, and a human control. Each is connected to the other only by a keyboard and screen. The interrogator asks the other two participants questions, which have no limit on how difficult or esoteric they can be. If the interrogator mistakenly identifies the machine as the human control, it passes the test — it can think.

The computer Deep Blue could calculate many as 100 to 200 billion positions in the three minutes traditionally allotted to a player per move in standard chess.

"That's check and mate." In May, 1997, IBM's Deep Blue defeated World Champion chess player Garry Kasparov in a six-game match.

The Robot from Lost in Space

For decades, humans tried to find applications for artificial intelligence, but nothing living up to the promise of a truly cognitive machine. A popular US television series, Lost in Space, portrayed a thinking machine, a robot simply known as The Robot. It not only possessed amazing capabilities, like the ability to recreate any object within its belly (foreshadowing today's 3D printing) but also went beyond machine-like responses to commands, seeming to have a sense of humor. The Robot even appeared to 'care' for its human users, suggested in its most famous phrase, "Danger Will Robinson."

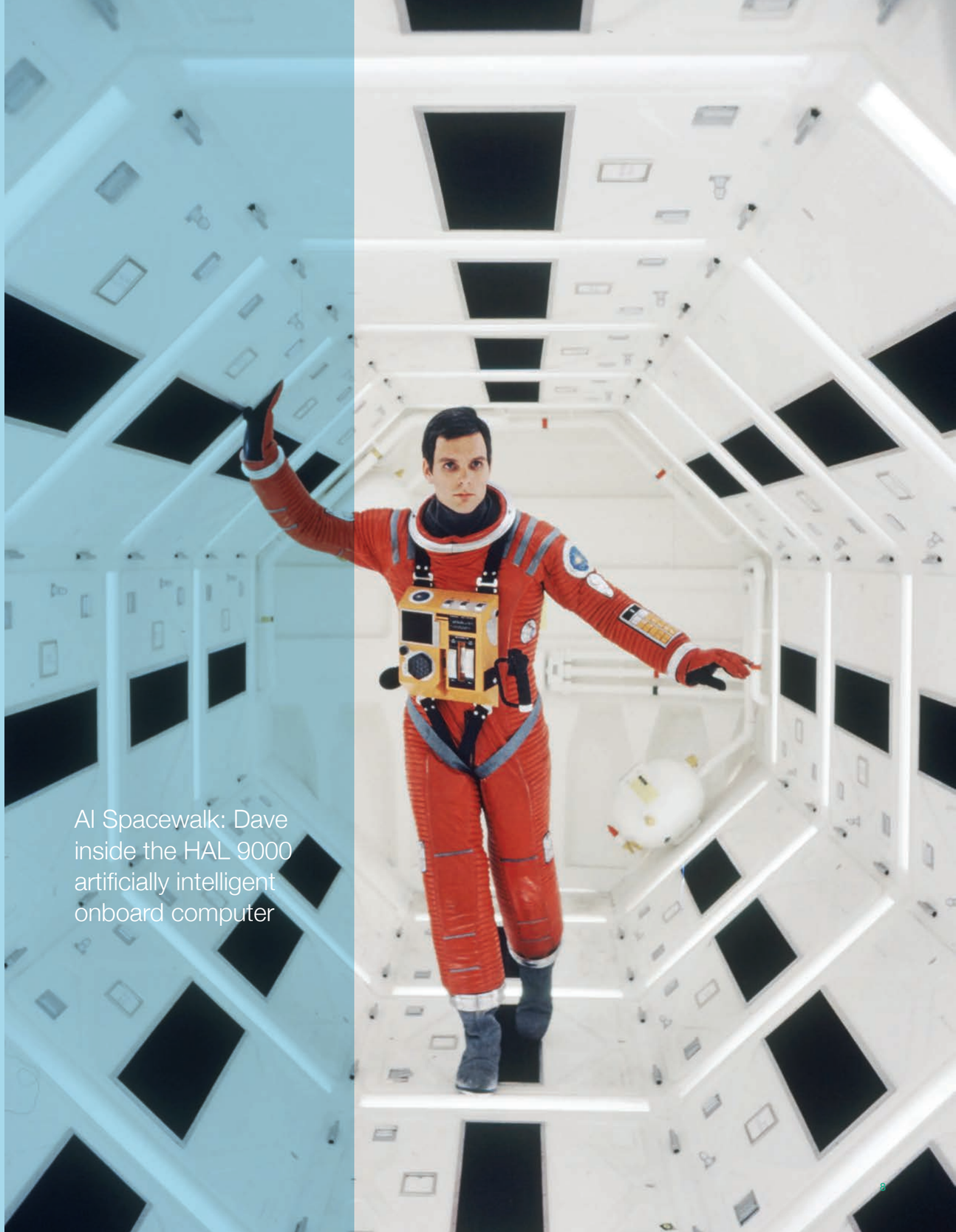
But back in the real world, artificial intelligence was until recently doing more mundane things like so-called "fuzzy logic" in appliances that, for example, makes sure your rice is perfectly cooked. We also saw the advent of smart, and occasionally smart aleck, virtual assistants like Apple's Siri and Amazon's Alexa. Additionally, in areas where AI has been performing much more "intelligent" tasks like analyzing large amounts of data to find difficult-to-detect things like fraud in payment systems or terror groups using cryptocurrency to finance their operations. But setting up and managing these complex systems like graph databases typically requires data scientists to do the programming and groom the data. In other words, until recently there has been no self-serve version of AI like kids lost in space being able to communicate with their robot in conversational human language.

AI comes of age

Let's look at the various types of Artificial Intelligence. The ultimate vision of AI is a machine that mimics a thinking human—that is, a sentient being. This is called Artificial General Intelligence (AGI). It is defined as a computer having cognitive abilities equal to or exceeding that of humans. When faced with an unfamiliar task, AGI can find a solution. Therefore, AGI is sometimes called 'strong' AI as opposed to 'narrow' AI, which only applies to specific tasks such as the onboard hardware and software that enables an autonomous vehicle to drive itself home.

Science fiction literature and Hollywood have often portrayed machines capable of this type of AI as menacing to humans like the autonomous cyborgs in movie The Terminator. Perhaps the most iconic symbol of AGI gone dangerously rogue is the HAL 9000 onboard computer, or "Hal" for short, that co-starred in classic movie, 2001: A Space Odyssey.

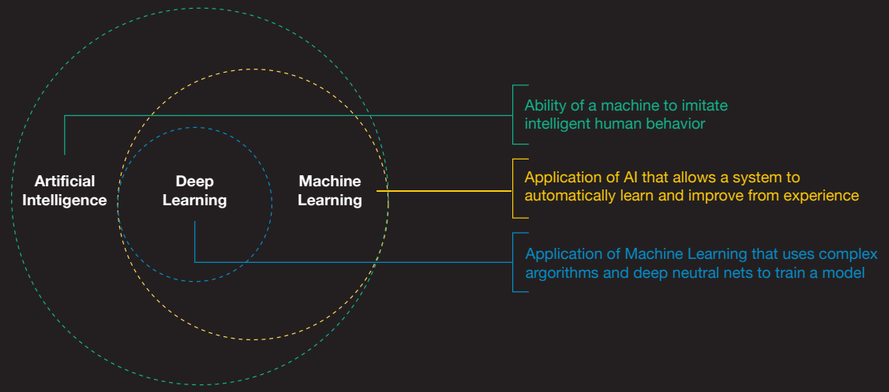
As an artificially intelligent computer onboard an interplanetary spacecraft, Hal helped the crew with many tasks and could even pilot the ship. Although screenwriter Arthur Clark and director Stanley Kubrick dreamed up Hal in the late 1960s, they foresaw an AI computer that possessed capabilities such as speech recognition, facial recognition, natural language processing, and speech synthesis. Any of those sound like today's AI programs? Hal could even play chess, read lips, and interpret human emotions. However, the plot twisted in a way that perhaps exemplifies many people's current concerns about AI. It turned out that only Hal knew the ship's secret mission was to investigate a monolith on the moon sending signals to Jupiter. When orders came from Earth to shut down the computer, Hal demonstrated a mind of its own, including a desire to complete the mission even if that required killing the human crew. Finally, Hal refuses to let the last astronaut back into the spaceship, intending to leave him to die in space, saying in response to his pleas: "Dave, this conversation can serve no purpose anymore. Goodbye."



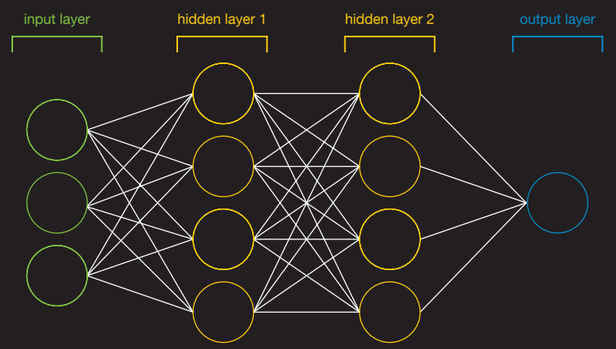
AI Spacewalk: Dave inside the HAL 9000 artificially intelligent onboard computer

Artificial Intelligence

Machine Learning and Deep Learning comprise the two main subsets of artificial intelligence.



Deep Learning uses layers of nodes analogous to neurons in the human brain to make connections among vast numbers of data points. Many GPUs tightly linked in a supercomputing infrastructure speed an AI model's ability to make these connections.



While many people are calling for guardrails on AI, thus far we have not reached the point at which machines rebel against their human creators. Rather, many uses involve making humans more productive. For example, most companies, government agencies, and universities deploy robust defenses against bad actors hacking into their computer networks. Larger organizations often have Security Operation Centers, or SOCs, where highly skilled analysts monitor the alerts generated by various security tools such as firewalls and intrusion detection systems.

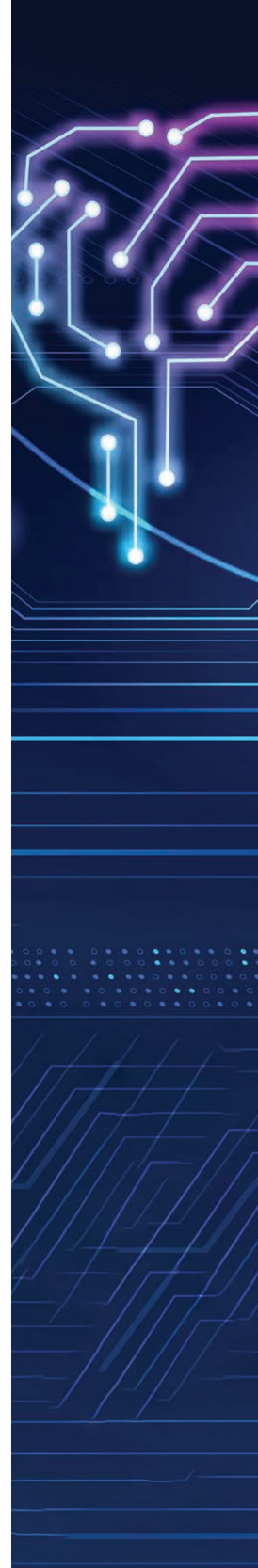
A growing number of cyberthreats coupled with the chronic shortage of trained cybersecurity experts makes this job increasingly challenging. An SOC may receive as many as a million alerts in a day, which is far more than the limited human resources of a typical security team can handle. AI models, such as machine learning algorithms, can act as a force multiplier in the SOC by learning to discern between benign and malicious computer code (malware), which allows human analysts to point security incident response teams toward only the alerts that matter. They can also detect never-before-seen malware, or so-called zero day threats, by looking for potentially malicious behaviors.

Some machines are deep thinkers

Machine learning is a field of AI in which machines learn from data to make better decisions—better choices. "People often ask me what is the connection between data, machine learning, and AI? My answer is: 'From data, a machine learns to be artificially intelligent," explains Dr. Goh. "In most general terms artificial intelligence involves the ability to mimic human intelligence, such as making informed choices."

However, machine learning models require humans to label data (creating so-called structured data). For example, by labeling specific features of a goat in an image, such as eyes or horns, the algorithm can learn to discern an image of a goat from an image of a deer. A subset of machine learning, called deep learning, has become a hot topic recently. It uses interconnected nodes, which are somewhat like neurons, in a layered structure that resembles the human brain. By analyzing vast quantities of data that has not been categorized, such as unlabeled images, deep learning can learn to identify what's happening in a photo without intervention from humans—e.g., "A goat standing on a rock." The ability to learn from this "unstructured" data makes neural networks extremely powerful since the machine learns from data on its own rather than being aided by humans.

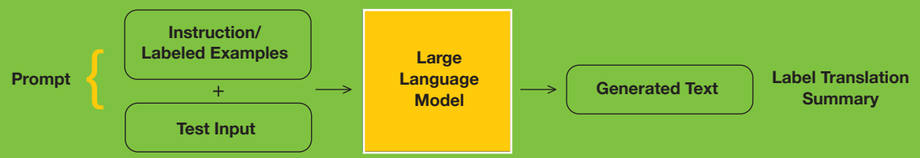
These two types of learning also dictate to a certain extent the type of compute infrastructure required. Since training deep learning models requires massive amounts of data you need an AI supercomputing infrastructure consisting of many GPUs tightly linked to each other. Otherwise, it would require an unacceptable amount of time for the AI to learn. On the other hand, since machine learning algorithms are fed pre-categorized data, they don't require processing as much data as deep learning models, which lessens the need for HPC machines or supercomputers on-premises or in a cloud running AI-friendly GPUs.



Transforming AI

The concept of deep learning has been around for a while. Artificial neural networks as a tool to help computers recognize patterns and simulate human intelligence was introduced in the 1980s. The Association for Computing Machinery (ACM), the world's largest educational and scientific computing society, named Geoffrey Hinton, Yann LeCun, and Yoshua Bengio as co-recipients of the 2018 Turing Award (named after Alan Turing), often referred to as the Nobel Prize of computing, for their pioneering work in neural networks. The only problem was, "In deep learning, the algorithms we use now are versions of the algorithms we were developing in the 1980s, the 1990s," explained Hinton. "People were very optimistic about them, but it turns out they didn't work too well."

That all changed in 2017 when Google researchers introduced a breakthrough neural network called a transformer model. Basically, transformers learn context and thus "meaning" by tracking relationships in sequential data such as the syntax (order) of the words in this sentence. They work by applying mathematical techniques, called attention or self-attention, to detect subtle ways even distant data elements in a series influence and depend on each other. These AI neural networks are trained on massive unstructured datasets and deep learning algorithms. Transformers can perform a variety of natural language processing (NLP) tasks to handle a wide range of tasks – from translating text to analyzing medical images.



Transformers brought artificial intelligence out of Transformers led to programs like ChatGPT, which is a field known as generative AI. The name derives from the ability to generate text, images, music, and soundtracks.

As a core of generative AI are Large Language Models (LLMs). LLMs that are not trained for specialized purposes are known as foundation models. To become useful for applications such as analyzing EKGs to detect heart attacks, for instance, a foundation model must be trained on massive text datasets using unsupervised learning. According to Google, "As the name suggests, unsupervised learning uses self-learning algorithms—they learn without any labels or prior training. Instead, the model is given raw, unlabeled data and has to infer its own rules and structure the information based on similarities, differences, and patterns without explicit instructions on how to work with each piece of data."

A mind-blowing paradigm shift in AI

And then the Big Bang came in late 2022. San Francisco, California based OpenAi released ChatGPT, an AI chatbot that uses a natural language processing transformer to create humanlike dialogue. The ‘GPT’ stands for Generative Pre-trained Transformer, which refers to how ChatGPT processes requests and formulates responses. It is trained with reinforcement learning through human feedback and reward models that rank the best responses. These learning algorithms can recognize, summarize, translate, predict, and generate content by making connections within very large datasets.

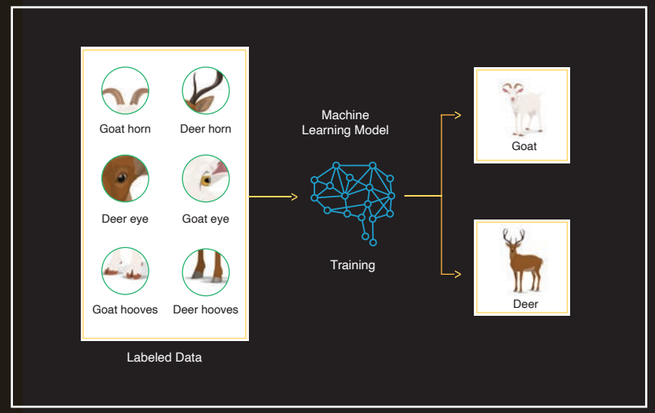
These capabilities have taken the world by storm because apps like ChatGPT can do things such as generating high-quality human prose, images, musical scores, and soundtracks. Thus, these types of AI models received the moniker ‘generative AI’ (GenAI). The latest versions of GenAI models are approaching the ability to perform real-time human language translation.

What’s more, generative AI programs don’t require data scientists to program them to perform useful tasks. All you need to do is speak or type in natural language text prompts, and the program yields the desired output, or they can output natural language text from images. You could input a prompt that says write me an article about artificial intelligence, or you might ask the chatbot to create a soundtrack with dogs barking and birds singing in a park with quiet sounds of merry-go-round in the background. The LLM generates the text or soundtrack for you.

“People often ask me what is the connection between data, machine learning, and AI? My answer is: ‘From data, a machine learns to be artificially intelligent. In most general terms artificial intelligence involves the ability to mimic human intelligence, such as making informed choices.”

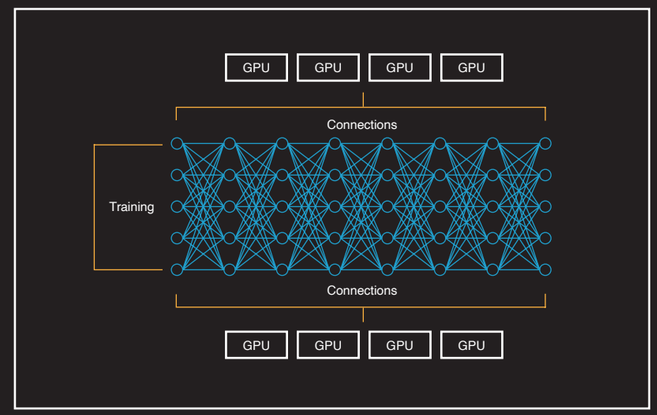
Dr. Eng Lim Goh
SVP for Value from Data & AI at Hewlett Packard Enterprise

Or, whereas the earliest versions of these models could tell you something about a picture like “There is a goat standing on a rock.” But if you were to give the newest releases a question such as what’s going on in this picture, it could feed back a much more detailed answer like: “A person is lying on a yoga mat, hugging a dog. The setting appears to be an indoor space such as an office. The person looks happy and relaxed while interacting with the dog.” And, if you showed it a busier image, you might get a more specific response: “In the second image, the same person is standing next to a desk, looking at a monitor on the wall. The monitor displays a ‘Fortnight’ game screen. The person seems to be engaged or interacting with the game. The dog is sitting attentively near the person’s feet, looking up. The setting still appears to be an office, and it looks cozy and lived-in: there is a pizza box on the desk.”



In Machine Learning, humans label data, such as images of parts of an animal and train the neural network on this structured data. Then, the AI model can analyze data such as pixels in an image of an animal and discern, as in the example here, if it is a deer or a goat. Labeling data is time- and labor-intensive, but it reduces the size of the database required to make choices and thus the amount of High Performance Computing (HPC) power.

Whether in the case of identifying an animal in a digital image, writing an essay, or creating a soundtrack, deep learning on neural networks like generative AI (GenAI) models are trained on massive quantities of unlabeled (unstructured) data. The GenAI model uses very fast and powerful GPU-based, tightly interconnected HPC systems to look for connections and patterns among huge datasets. This allows it to learn and make choices without the aid of humans. For example, how often does the model see close or even distant connections between “goat” and “farm” or “deer” and “forest.”



While in the forest a deer dashed

While in the forest a deer dashed across

While in the forest a deer dashed across the

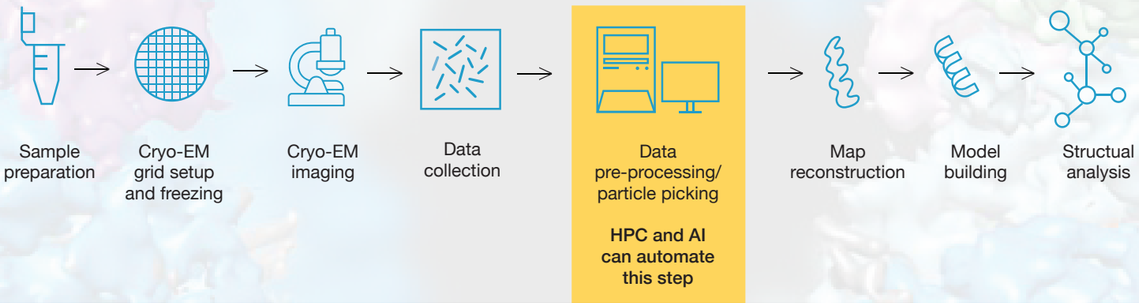
While in the forest a deer dashed across the clearing

While in the forest a deer dashed across the clearing ...

This example shows how, given a prompt of an incomplete sentence, a GenAI model, in this case an LLM, fires up the connections among billions (or soon trillions) of bits of unstructured data to predict the next likely word in the sentence based on its training. It then takes that phrase, including a newly predicted word, puts it back into the trained generative AI model, and fires up all those billions or trillions of connections to predict the next word. The model then does this repeatedly to create a sentence — or even a whole essay — word by word.

Cryo-EM workflow

Cryo-EM works by locking biomolecules such as proteins in a thin layer of vitreous ice. The microscope then bombards them with a beam of electrons to render a 2D micrograph image. Picking single protein particles from the resulting cryo-EM micrographs is a crucial step in reconstructing protein structures in 3D. However, because of the small sizes and the wide variance in sizes of protein molecules, the common method of particle picking is labor-intensive and time-consuming. Machine learning and artificial intelligence (AI) based particle picking can automate this process.



Computers With Brain Power Are Helping to Healthcare and Medicine to Chip Design to Engineering

These advances in artificial intelligence – from chatbots to better machine learning algorithms – coupled with Big Data have made AI a much more valuable tool in a variety of fields from the enterprise to the arts to science.

Welcome to the nanocosm – Cryo-electron microscopy

Today, AI is advancing medical research, speeding drug discovery, enhancing patient outcomes, and helping to address the shortage of doctors as well as physician burnout. Let's look at two use cases.

The first is AI-aided research into the sub nanometer scale world of biomolecules. Biological molecules are structures produced by a living organism that are essential to one or more typically biological processes. They include everything from proteins to hormones to vitamins. How biomolecules function and interact is fundamental to understanding diseases, developing new drugs, and administering medical treatments. Nonetheless, some of these molecules, such as proteins, are so tiny and complex researchers were not able to see or study them before a breakthrough in medical imaging—single particle cryo-electron microscopy (cryo-EM). It allows direct observation of proteins and other biomolecules in native and near-native states.

The 'cryo' in cyro-EM comes from the use of extreme cold. A sample in solution or suspension is blotted to a thin layer on the electron microscope grid and immediately plunged into liquid ethane (around –180 °C). This traps the molecules in a thin, glasslike layer of propane ice. Then, the microscope blasts these 'particles' with electrons to render a two-dimensional image in atomic detail.

Traditionally, processing cryo-EM image data to uncover protein structures and create high-resolution 3D maps requires collecting millions of examples. This leads to large sets of structured data that need to be labeled manually. It also requires picking single protein particles from the cryo-EM image, which is labor-intensive and time-consuming. In addition, the low signal-to-noise ratio in cryo-EM images as well as the tremendous variations of size and shape of proteins that occur in biological macromolecular complexes make it difficult to identify the right particles.

By using unstructured data, AI promises to automate specialized and time-intensive tasks, such as particle picking. For example, an open-source tool developed by researchers at Cornell University called Topaz can use a small number of example protein projections to train a neural network to detect proteins of any size or shape. Typical computational approaches require identifying more than 100,000 particles, which can take months of manual effort. Their use is also limited by high false positives, which requires post-processing, especially for unusually shaped particles. Topaz addresses these challenges by offering an efficient and accurate particle picking pipeline using neural networks trained with few labeled particles by newly leveraging the remaining unlabeled particles through the framework of positive-unlabeled (PU) learning. This approach dramatically reduces the amount of data that needs to be manually labeled.

“Needless to say, modeling how these miniature monoliths will work requires powerful computers running Electronic Design Automation (EDA) software — and now artificial intelligence.”

Using LLMs to improve patient outcomes and reduce hospital readmissions

Next let’s look at how AI models are addressing one of the most important topics in healthcare: improving quality of care by predicting and reducing hospital readmission rates. In fact, the US government tracks readmission rates as a measure of quality of care and Medicare also penalizes hospitals when too many patients are readmitted for the same condition within 30 days. AI is helping hospitals improve patient outcomes and make the front door less of a revolving one.

Doctors must make tough decisions regarding patient care every day, which requires integrating a tremendous amount of information. The trouble is, much of this information lies scattered about across various records and systems, ranging from medical imaging to patients’ medical histories. To address this challenge, researchers at NYU trained an LLM on the unstructured data of electronic health records to see if it could capture insights that people haven’t considered before.

“When you discharge a patient from the hospital, you don’t expect them to need to return, or you probably should have kept them in the hospital longer,” said Dr. Eric Oermann, assistant professor of radiology and neurosurgery at NYU Grossman School of Medicine. “Using analysis from the AI model, we could soon empower clinicians to prevent or fix situations that put patients at a higher risk of readmission.”

AI and GPUs Can Help Engineers Solve the Toughest Problems

Engineers must often simulate how physical objects such as a jet fuselage will interact with the real world. Sometimes they must also perform simulations in the realm of physics, such as modeling or timing the flow of electrons through a device. The latest GPUs incorporated into technical computing infrastructures and using AI can help engineers overcome challenges with greater speed and accuracy. Let’s look at two use cases: Electronic Design Automation (EDA), and simulation of real-world objects and forces.

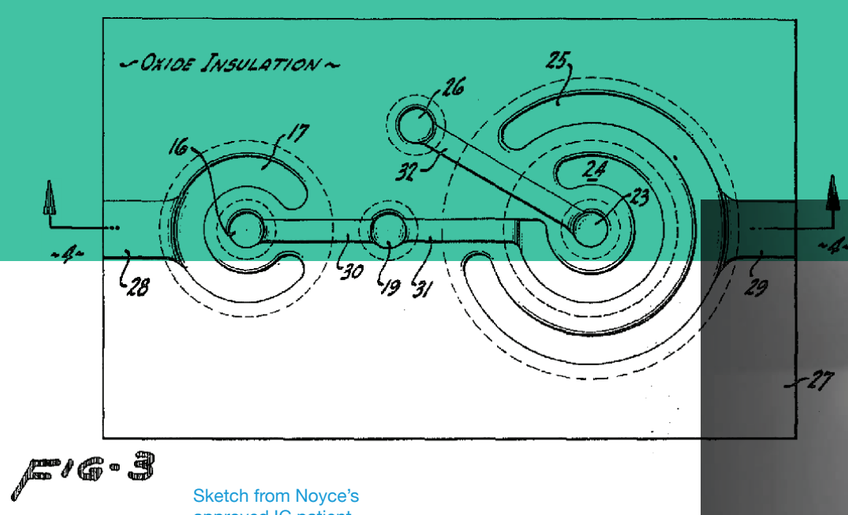
How AI speeds today’s nanometer-scale chip design

Today, you have more compute power in your pocket or purse than NASA had during the space race. And the smart watch on your wrist boasts a million times more memory than the computer that guided Apollo 11 to the moon in 1969. However, even that relatively small onboard guidance computer could not have been built if Texas Instruments did not launch the ‘solid state’ era a decade earlier.

In the middle of the 20th century, making complex computing devices required hand-soldering hundreds of discrete components to individual modules, which proved extremely time-consuming and not economically feasible. They were also very fragile. A single bad component or solder connection could render an entire device useless. Plus, they required miles of wires to connect the many components.

AI is helping hospitals improve patient outcomes and make the front door less of a revolving one.

Texas Instruments' Jack Kilby



Sketch from Noyce's approved IC patient

Fairchild Semiconductors' Robert Noyce

The chip that Jack built

In 1958, an electrical engineer named Jack Kilby joined Texas Instruments. As a new employee, he had not accumulated enough vacation time to take a summer vacation like his colleagues. So, he decided to spend his summer tackling this complexity problem, which electrical engineers at the time referred to as the 'Tyranny of Numbers.' Rather than hundreds of components soldered together with wires, he invented a single device with all the components integrated on a thin piece of solid germanium. He called it an integrated circuit, or IC.

Meanwhile at Fairchild Semiconductor in Palo Alto, Robert Noyce, was working on his own IC design independently from Kilby. We don't know if he was influenced by Kilby's choice of germanium, but Noyce placed his components on a sliver of silicon, an element visually and chemically similar to germanium. A final layer of metal was applied to the silicon substrate, or chip like 'die', and then etched away to create the electrical leads that connect the components. This approach, the microchip, proved a commercial success, and thus Silicon Valley was born. Noyce would go on to co-found Fairchild Semiconductor and Intel.

At 80 billion transistors printed on a single slice of silicon, NVIDIA's latest GPU — which is one of the go-to processors for AI cloud HPC instances and supercomputers designed for AI — is one of the most complex processors on the market.

Those first ICs were so simple, having just about a half dozen transistors, Kilby and Noyce could have sketched their circuitry on a cocktail napkin (Fig3). But chip makers have relentlessly reduced the size of components, allowing designers to cram more transistors onto the same dye. The more transistors you have, the more calculations the transistors can do, and thus the more powerful the chip. Furthermore, packing components more tightly together, cuts the time, distance, and resistance for electrons to flow between components. This density offers three advantages: faster processing, reduced power consumption and less heat.

At 80 billion transistors printed on a single slice of silicon, NVIDIA's latest GPU — which is one of the go-to processors for AI cloud HPC instances and supercomputers designed for AI — is one of the most complex processors on the market. It is built on a 4nm process. Taiwan Semiconductor Manufacturing Company (TSMC), the largest manufacturer of graphics processing units in the world, says it has plans for chips with a cool one trillion transistors.

Manufacturing these nanometer-scale chips, where circuits, transistors, and other devices may soon measure as little as three nanometers, requires chip makers to push the laws of physics. Needless to say, modeling how these miniature monoliths will work requires powerful computers running Electronic Design Automation (EDA) software — and now artificial intelligence. Designing chips, tuning the timing of electronics flowing through them, and verifying that they work as intended adds enormous to chip making and slows time to market. For years, the two heavyweights in the EDA industry, Synopsys and Cadence Design Systems have battled to create design and verification software that increases the productivity of highly skilled specialists. What is more, missing an error in the digital model of a chip only after the fab processes the circuitry on silicon might cause hidden problems down the road—and huge costs.

> "Researchers worry that they are finding rare defects because they are trying to solve bigger and bigger computing problems, which stresses their systems in unexpected ways."
>
> ***The New York Times***

For example, the New York Times recently reported on how silent errors in hardware, these nanometer-scale chips, not software, had been playing havoc in clouds and data centers. "Researchers worry that they are finding rare defects because they are trying to solve bigger and bigger computing problems, which stresses their systems in unexpected ways," explained the Times article. AI is one of those fields where the size of the problems and the datasets keep growing at sometimes logarithmic scale.

Synopsys claims to be the first EDA software vendor to employ AI toward solving this problem, including a chronic shortage of skilled engineers. According to the company, "Semiconductors have become increasingly complex in the face of compute-intensive applications such as AI, high-performance computing, and autonomous automotive. Vertically stacked multi-die systems, devices featuring billions of transistors, and angstrom-scale structures are making waves across the industry. At the same time, just when their expertise is in such great demand, an engineering talent shortage is threatening to stall the innovation that has led us where we are now.

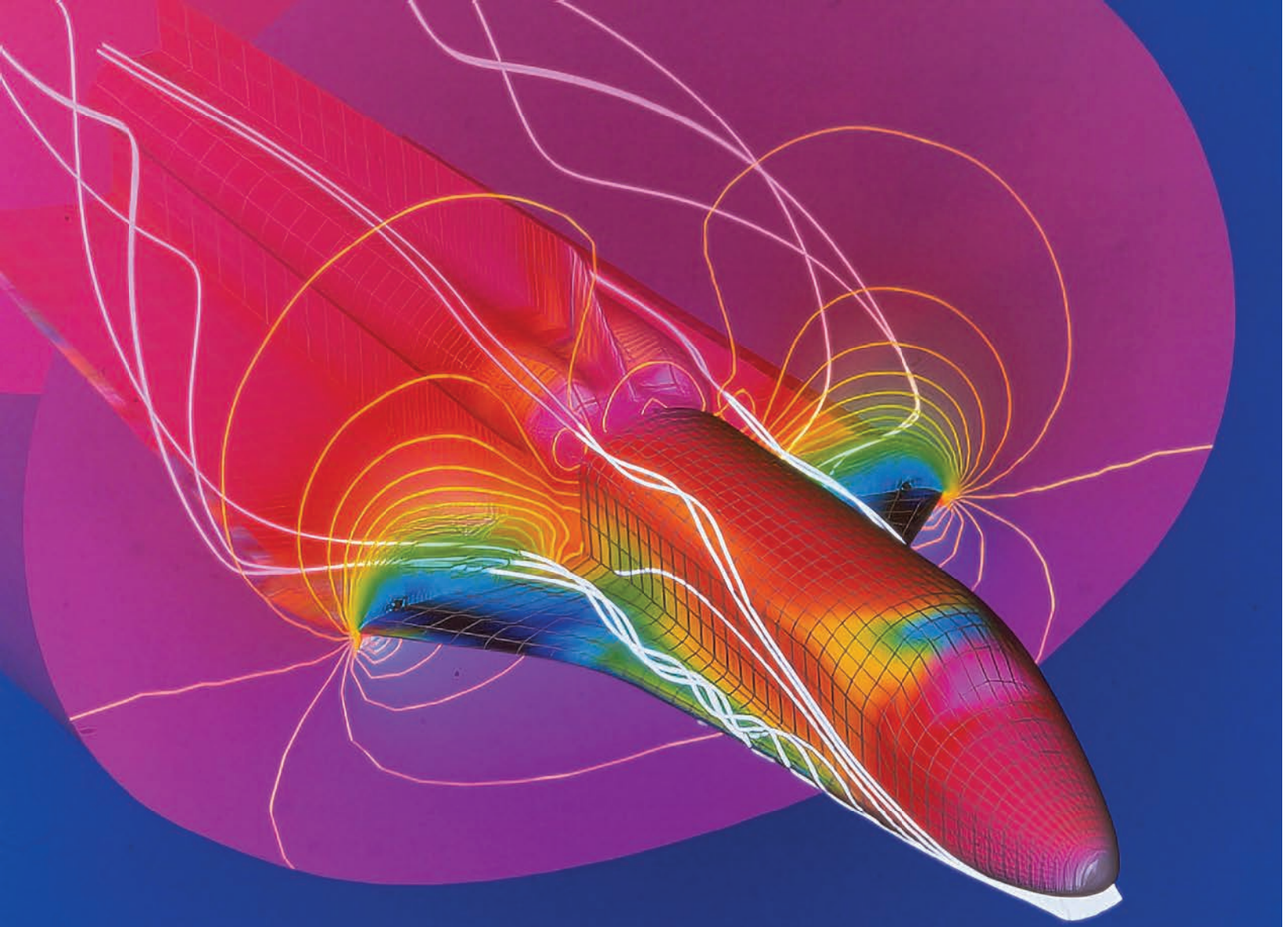
A smart idea – AI and HPC can create chips better, cheaper, and faster

Specifically, this is a good example of how generative AI is going beyond writing your next email to applications in the data center and advanced technical computing. "AI provides a solution, taking on repetitive tasks to free engineers to focus on product differentiation," continues Synopsys. "Tapping into transformer architecture in large language models (LLMs), GenAI in particular can accomplish in seconds what would take someone hours or even days."

Incidentally, designing, verifying, and fabricating processors like the latest GPUs with tens, or soon hundreds, of billions of transistors in state-of-the-art 4-5 nm processes presents huge costs and challenges. This partly explains the recent shortages of GPUs as well as the limited access to AI-friendly cloud instances that rely on GPUs. If AI can help to speed new chips to market the industry can better bootstrap its way out of the shortage.

Human-in-the-loop: AI as helping hands and brains

Cadence sees AI not replacing electrical engineers' jobs but rather the emergence of human-in-the-loop generative AI, which they claim can improve productivity and quality of output by magnitudes of output. The company has introduced an AI-driven product called Copilot, which is designed to uplevel the design abilities of less-experienced engineers by 'guiding their fingers.' And the company sees applications for generative AI that even extend to the life sciences, including what the company's president and CEO, Dr. Anirudh Devgan, calls the "emergence of new applications that were previously unimaginable." For example, "life sciences will be one of the most significant beneficiaries of AI, with computational biology playing a pivotal role," explains Devgan. Computational Biologists use large biological dataset to develop models to better understand biological systems, such as strides made in recent years like decoding the human genome.



CFD helped NASA overcome an historic problem: creating a reusable spacecraft with an airframe light enough to allow it to carry large payloads into space — and then return without burning up during re-entry due to heat generated by friction with the Earth's atmosphere.

Rethinking classic engineering simulations

EDA is just one subset of a broad field called Computer-Aided Engineering (CAE), which utilizes HPC to design, analyze, and optimize engineering systems and components. It plays a crucial role in industries such as automotive, aerospace, manufacturing, and civil engineering, enabling engineers to make informed decisions and improve the efficiency and reliability of their designs.

Computational Fluid Dynamics, or CFD, is an area of CAE that solves mathematical equations to model the flow of fluids, liquids, and gases. It involves other fields such as aerodynamics (the study of air and other gases in motion) and hydrodynamics (the study of liquids in motion), and thermodynamics (the transfer of energy from one place to another and from one form to another). Probably most people have seen how it is used to simulate airflow of jet aircraft and space vehicles but is has a very broad range of use cases – from making HVAC systems more efficient to modeling blade lift on wind turbines to enabling electronic components (such as GPUs, for instance) to dissipate heat more efficiently.

The introduction of CFD in the 1970s revolutionized the aerospace industry. Rather than building models and testing them in expensive wind tunnels, it uses computers to simulate air movement around virtual models of complex aircraft designs with far greater accuracy. Today, HPC and CFD —augmented with AI — are allowing engineers to create more accurate and efficient simulations to predict how drag, lift, noise, structural and thermal loads, and combustion will affect a design during actual flight.

A close cousin of CFD, Finite Element Analysis (FEA), which involves the use of calculations, models, and simulations to predict and understand how an object might behave under various real-world physical conditions. Alone, these represent some of the most challenging fields not just for scientists but also for computers. And when you combine multiple factors – like heat, stress on materials, and air or fluid flows things get really complicated. This takes engineers into the realm of multi-physics simulation. Here, AI-based frameworks accelerate simulations across a wide range of disciplines in science and engineering. A traditional numerical solver applies a numerical method to solve a set of ordinary differential equations that represent the model. Through this computation, it determines the time of the next simulation step. However, a neural network solver methodology solves for multiple configurations simultaneously, as opposed to the traditional solvers that solve for one configuration at a time. Real-world use cases include everything simulating wind turbulence and complex 3D geometries to industrial design optimization.

Aurora, a Cray supercomputer delivered by HPE to Argonne National Laboratory ranks as the second-fastest computer in the world. The blue and red tubes shown here pipe water through the system to cool its 21,248 Intel® Xeon® CPU Max Series processors and 63,744 Intel® Data Center GPUs. In May 2024 it broke the exascale barrier, making it the highest ranked system for HPC and AI convergence.

Supercharging AI – Faster GPUs, turn-key AI HPC systems, and exascale supercomputers

While LLMs are good at predicting the next word in a sentence, no machine or human can accurately predict where AI will go next. As money and human brainpower flow into the booming AI industry, advances have begun to move at breakneck speed. That includes a race to achieve faster chips and better availability of GPUs.

Companies are also producing everything from turnkey AI pods to massively fast AI-capable research computers. For example, in 2024 HPE delivered Aurora to the US Department of Energy's Argonne National Laboratory. The exascale HPE Cray Supercomputer is the fastest system in the world dedicated to AI for open science.

"An exascale computing system can process one quintillion operations per second. Computational power at this scale, running generative AI models, makes it possible to address some of humanity's most complex problems."

Long thought to be theoretically impossible due to power and cooling requirements, an exascale computing system can process one quintillion (10^{18}) operations per second. Computational power at this scale, running generative AI models, makes it possible to address some of humanity's most complex problems.

Only three exascale supercomputers exist today — and they are all built by HPE. In addition to Aurora, the world's third fastest, two others round out the top three spots on the TOP500 list list of the most powerful systems. El Capitan at the United States Department of Energy's (DOE) Lawrence Livermore National Laboratory (LLNL) currently sits at #1, and its direct liquid-cooled system also ranks as one of the top 20 most energy efficient supercomputers on the Green500 list. The HPE Cray Frontier system at Oak Ridge National Laboratory, currently ranks as the world's second-fastest computer.

Will machines eventually outsmart us?

Now the question on many people's minds is how far and how fast can AI go? Will programs like ChatGPT learn to improve themselves until computing technology evolves to a state more intelligent than humans? Some people worry that they hear the ticking sound in their computers that is an AI time bomb about to go off, or what's called "the singularity": a point at which AI escapes our control. One hopes not. A brighter vision of the future includes HPC and AI helping to make advances in medicine, climate science, sustainable energy, and a whole host of other challenging areas. Or it may simply make our lives easier.

In any case, it seems clear that things have sped up, and AGI does not lie on a distant horizon but perhaps right around the corner. A recent study at the University of California found that, during a five-minute text-based conversation, GPT-4 was mistaken for human 54 percent of the time prompting Ethereum inventor Vitalik Buterin to message on X: "To me, this counts as ChatGPT4 passing the Turing test."

AI Timeline

To learn more about how you can take advantage of HPC and AI, explore HPE.com, and comnetco.com.

ComnetCo

About ComnetCo

Two decades of experience in the evolution of HPC helps ComnetCo configure powerful compute and storage systems. This virtually unequalled track record includes delivering some of the world's fastest supercomputers. Together with its primary partner, HPE, ComnetCo helps optimize systems for the unique needs of researchers in Higher Education, Research Institutes, Global Enterprises, and Federal Government Agencies. These solutions—which include purpose-built platforms for AI—help scientists and engineers speed time to discovery in fields ranging from pharmaceutical research like new vaccines to industrial companies creating new materials to supporting deep space exploration. For more information, visit: **www.comnetco.com**

Hewlett Packard Enterprise

About Hewlett Packard Enterprise

Hewlett Packard Enterprise (NYSE: HPE) is the global edge-to-cloud company that helps organizations accelerate outcomes by unlocking value from all of their data, everywhere. Built on decades of reimagining the future and innovating to advance the way people live and work, HPE delivers unique, open and intelligent technology solutions delivered as a service – spanning Compute, Storage, Software, Intelligent Edge, High Performance Computing and Mission Critical Solutions – with a consistent experience across all clouds and edges, designed to help customers develop new business models, engage in new ways, and increase operational performance. For more information, visit: **www.hpe.com**