OAK RIDGE NATIONAL LABORATORY
REVIEW

AI is revolutionizing industries and advancing research at ORNL, where experts are balancing its potential with its risks. Credit: Laddy Fields/ORNL, U.S. Dept. of Energy

Turning AI into something we can trust

In the good old days of the early '90s, the internet promised to make the world and its then-five and a half billion people better.

By giving us equal access to the world's information, it was going to democratize knowledge. By giving us all the ability to make ourselves heard, it was going to democratize publishing. By connecting everyone, everywhere, it was going to keep us in touch with friends and family.

We would make better decisions, have higher-paying jobs, be better global citizens.

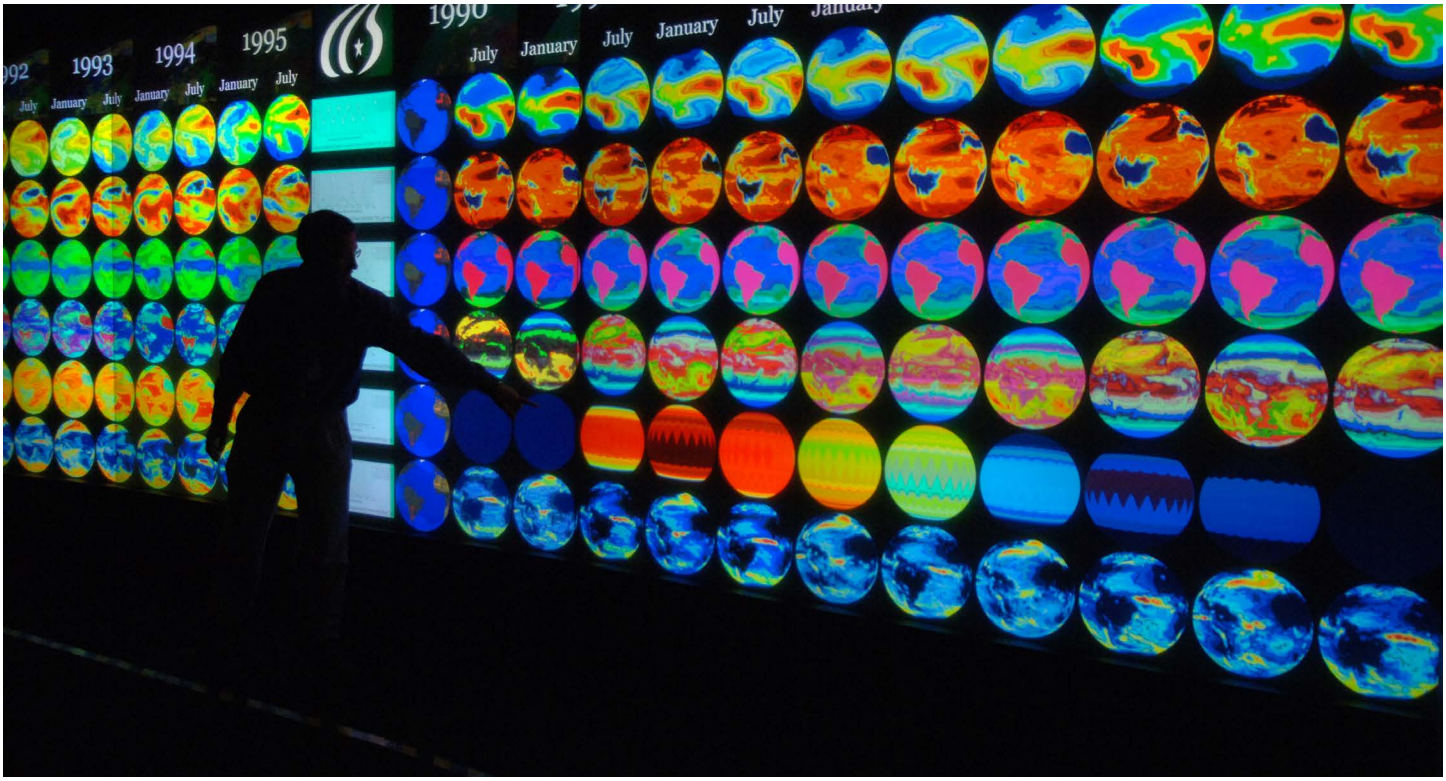
And the internet really did those things — to some extent. You can do your research without opening a book, track down long-lost friends, publish your thoughts for all to see.

But there's also been a dark side. While old friends can contact you with the touch of a button, so can spammers, scammers, stalkers and all kinds of people you'd rather not hear from. Click on the wrong link or trust the wrong message and you've given some stranger entry into your bank account or the ability to attack other networks using your computer.

So while we were learning fun new terms like "browser," "dot-com," "dial-up" and "e-mail," we would shortly be learning far less fun meanings for terms like "virus," "worm," and "Trojan horse" and new portmanteaus like "spyware," "adware" and "ransomware."

As a result, the online world of 2024 often feels less like a playground and more like a war zone.

"If you connect a computer to the internet, it will be attacked within the next few minutes," said Edmon Begoli, director of the **Center for Artificial Intelligence Security Research** at Oak Ridge National Laboratory. "That's just how things are. In AI, we're not there yet, thankfully, and that's why we're doing what we are doing to protect AI systems."



ORNL's Artificial Intelligence Initiative

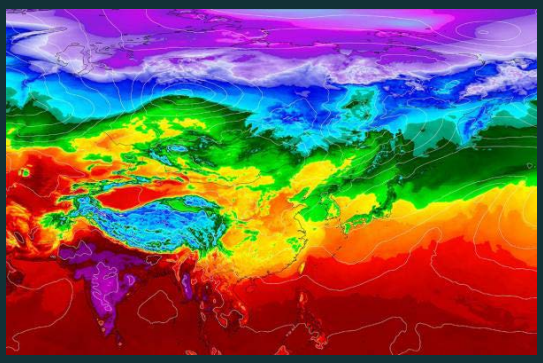
In a sense, artificial intelligence in the 2020s feels something like the internet of the 1990s: full of promise but soon to be full of menace. AI can be breathtakingly powerful, delivering everything from fast, accurate cancer diagnoses to cars that ferry you around without human involvement. But it can also be breathtakingly dangerous. The same tech that will produce miracle drugs can also produce terrible poisons. The same tech that ferries you around in driverless taxis is also controlling autonomous weapons.

"We have seen how AI can potentially transform scientific discovery and strengthen national security, but the potential benefits of AI are challenged," said Prasanna Balaprakash, director of **ORNL's Artificial Intelligence Initiative**. "They're challenged because of a lack of safety, security and trustworthiness, and AI models are energy-consuming."

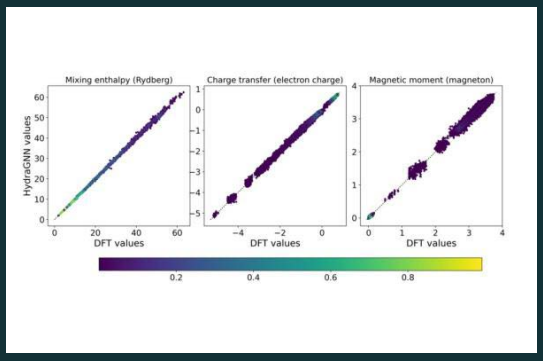
ORNL's AI Initiative is dedicated to the proposition that we can maximize the benefits of AI while mitigating its harms and, in doing so, make the world a safer place.

ORNL is, in fact, heavily invested in artificial intelligence. The lab is using AI for everything from controlling **multimillion-dollar scientific instruments** to **finding extremely rugged materials** to **discovering new drugs**.

Consider the following accomplishments by ORNL researchers:



Creation of the **largest-ever AI foundation model for climate science research**,the hundred billion-plus-parameter Oak Ridge Base Foundation Model for Earth System Predictability, or ORBIT. This model, a finalist for the prestigious **ACM Gordon Bell Prize for Climate Modelling**, is a thousand times larger than previous climate AI models.



Creation of an open-source architecture that will supercharge research into such diverse areas as drug design, nuclear shielding, materials characterization and electron microscopy. **Known as HydraGNN**, the architecture for graph neural networks — which process data that can be represented as graphs — is able to handle datasets comprising hundreds of millions of graphs.

A **study showing how a large language model might be most efficiently trained on a leading supercomputer**. The team used ORNL's world-leading Frontier supercomputer and test data to project how models with 22 billion, 275 billion and 1 trillion parameters could run on 128 and later 384 of Frontier's 9,400-plus nodes.

ORNL and AI go way back

ORNL's history with AI goes back at least to 1979, with creation of the **Oak Ridge Applied Artificial Intelligence Project**, a collaboration of mathematicians and scientists focused on evaluating AI's potential to bolster scientific research.

But AI as we know it didn't come into its own until the rise of the graphics processing unit, a chip dating back to the 1970s that was created to accelerate computer graphics. In the 2010s, GPUs became the driving force behind scientific computing as well as artificial intelligence.

ORNL was among the first computing powerhouses to go all in on GPUs with the 2012 introduction of the lab's Titan system. Titan debuted on top of the **Top500 list**, which ranks the world's fastest supercomputers.

The lab's next two supercomputers — Summit and Frontier — also debuted at the top of the list while focusing more directly on AI. Both are powered by GPUs; Frontier has 37,632 of them.

Frontier was also the first supercomputer to cross what is known as **the exascale barrier**. This means it can perform more than a quintillion calculations each second.

It's impossible to describe the speed of a modern supercomputer in any way that feels meaningful. To paraphrase Douglas Adams in "The Hitchhiker's Guide to the Galaxy," supercomputer speeds will not fit into the human imagination.

But here's a try: Frontier can perform 2 quintillion calculations in one second (that's a "2" followed by 18 zeroes). To accomplish the same feat at a rate of one calculation per second, it would take you more than 63 billion years. This is four and a half times the age of the universe.

So it's fast.

Secure, trustworthy, energy efficient

ORNL is well positioned to guide the future of AI. Not only is it the country's largest multidisciplinary national laboratory, it is also a dominant force in supercomputing and a major player in the United States' national security research community.

"What sets ORNL apart is how its AI efforts in science and national security enhance each other," Balaprakash said. "This synergy creates a unique and comprehensive AI program that not only advances important research but also provides practical and sustainable solutions to challenges in various fields."

This unique position guides ORNL's AI Initiative, which is pushing the boundaries of artificial intelligence for scientific discovery — tackling tsunamis of data to better understand and manage complex systems, and automating research facilities to make experiments more efficient and precise while freeing human researchers from repetitive tasks — while focusing on the triple challenge of security, trustworthiness and energy efficiency.

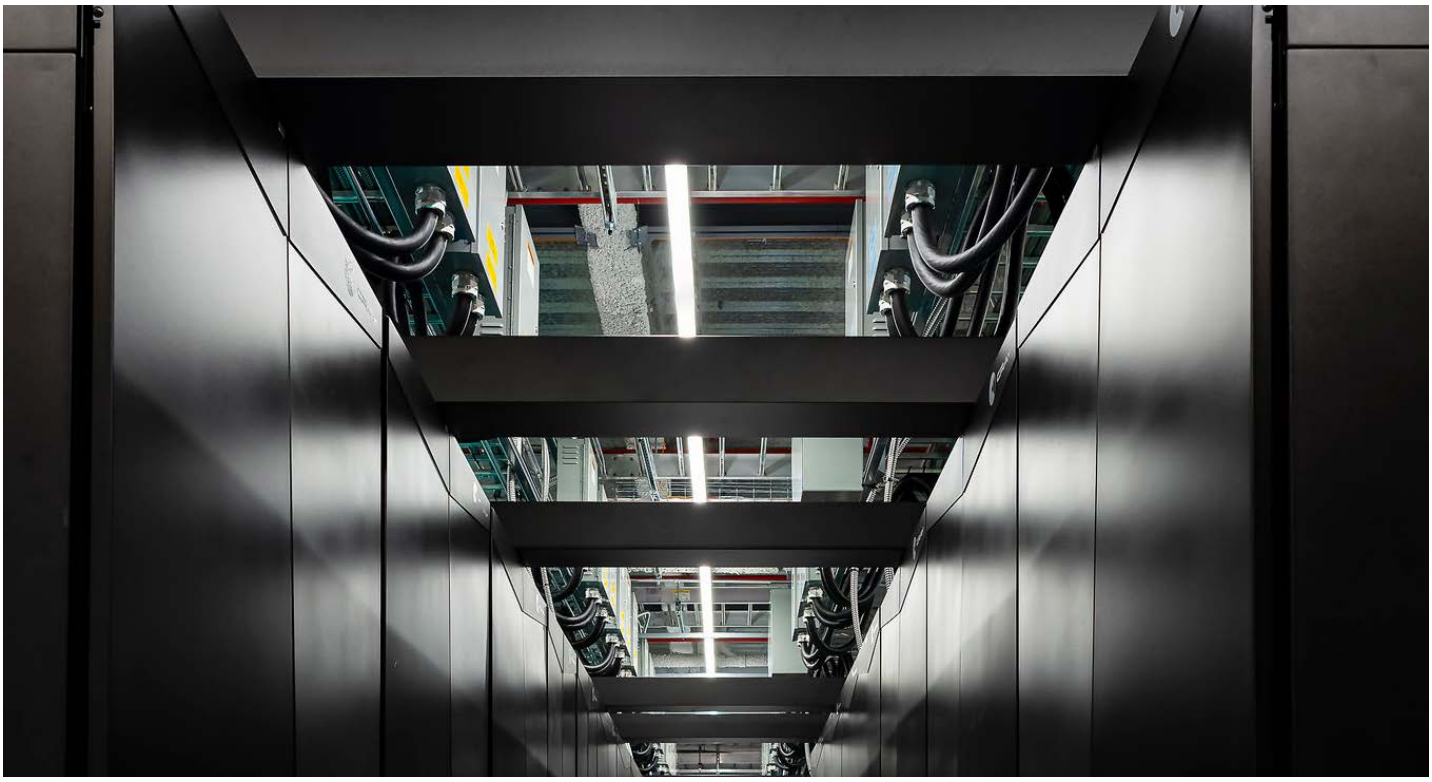
Security

In directing the **Center for Artificial Intelligence Security Research**, Begoli spends a lot of time considering the myriad ways artificial intelligence can be weaponized.

AI of course isn't the first tech to be misused by bad actors. Thieves and terrorists have seen the internet as a boon for decades. What distinguishes AI as a potential threat may be its surprising power.

Consider that in May 2024, Air Force Secretary Frank Kendall spent an hourlong flight in an F16 fighter that was completely autonomous, **piloted by AI**. This achievement followed **earlier exercises** in which an AI agent flew an F-16 in dogfights against a human crew. And that, in turn, followed simulations in which an AI agent beat a human pilot 5-0.

CAISER's, and Begoli's, goal is to identify ways to protect AI systems and the people affected by them before attacks become ubiquitous. In particular, the center focuses on end-to-end AI security, AI vulnerability research and AI security evaluations at scale, with current research looking at defenses against data poisoning, evasion attacks and the misuse of deepfakes.



Data poisoning

Many AI models are created to distinguish one thing from another: a cat from a dog, legitimate software from malware, a valid credit card transaction from a fraudulent one.

Classification may not be the most dramatic form of artificial intelligence, but it is one of the most ubiquitous — and important — jobs that AIs are called on to perform. And they have gotten good.

- When Visa examines transactions for potential fraud, **the company's AI looks at more than 500 different attributes for each**. Visa said it was able to block $40 billion in fraud between October 2022 and September 2023.
- NASA's Perseverance Rover **uses an AI technique** called adaptive sampling to **look for evidence of life on Mars**.
- A British collaboration **recently developed AI software** that can analyze chest X-rays for 37 separate conditions. In 35 of them, its analyses were at least as accurate as that of a human doctor.

Let's say you want your model to show people cat photos and only cat photos. To get there, you feed the model a hundred photos of cats — or a million — and it will get steadily better at distinguishing cats from not-cats. Show it a new photo of a cat, and it will identify it as a cat. Show it a new photo of a dog, and it will identify it as a not-cat.

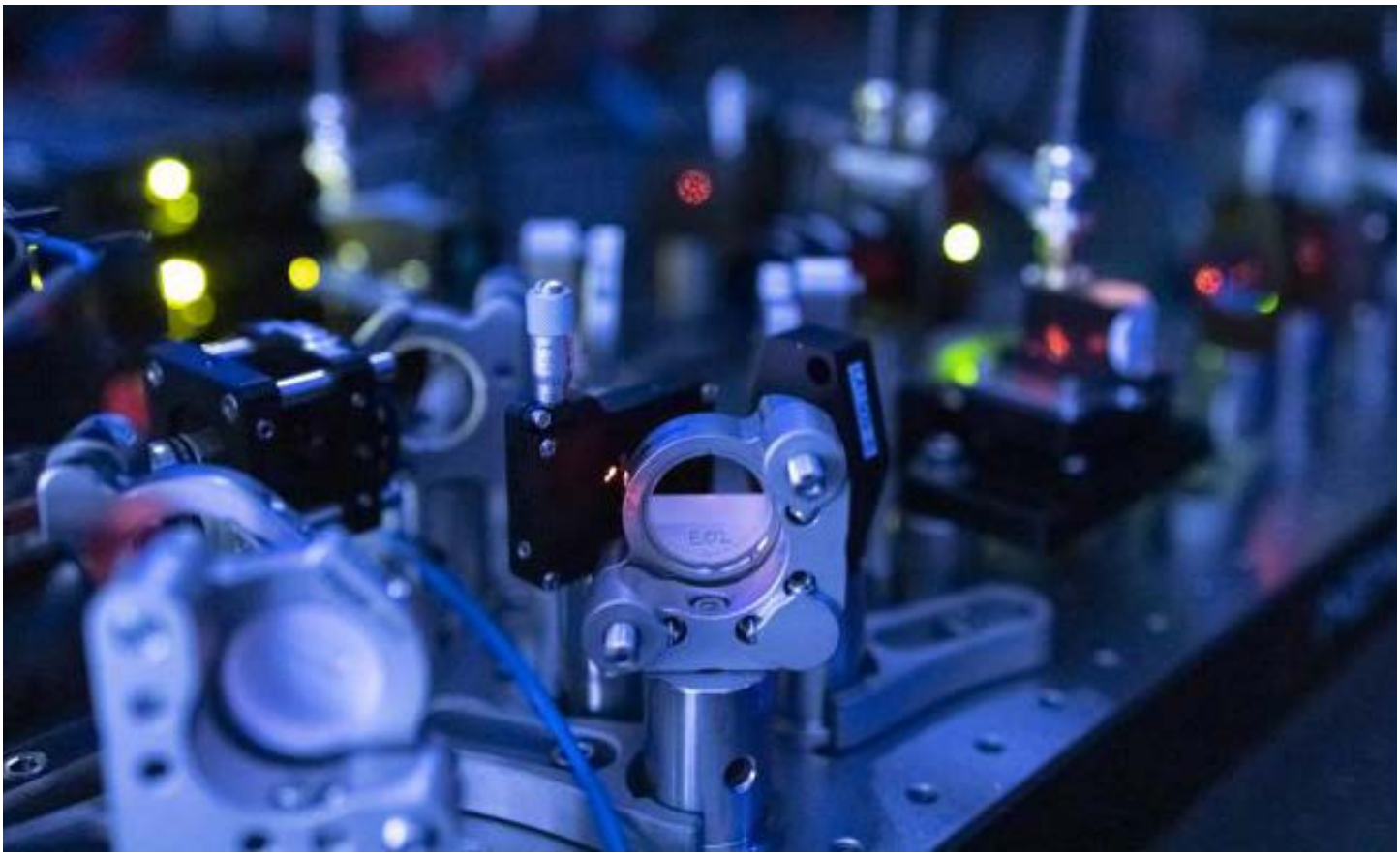
But what if you're a dog person and you want the cat model to mess up. That's easy enough to do — if you can act while the model is being trained. Simply feed it some dog photos and call them cat photos, and it will think at least some new dog photos are indeed cat photos.

Now let's say the AI is being trained to screen for malware, distinguishing legitimate software from malicious code.

Like cat photos, legitimate software has elements that an AI can use to identify it. Include malicious code in the AI's training set and you've created a backdoor that gets your malware through the screening process. Before you know it, people's personal and financial information is being stolen, ransomware is forcing them to pay ludicrous sums, and websites are being targeted by millions of computers, most of whose owners have no idea what they're participating in.

"That's data poisoning," Begoli said. "You're going into the training data. You're doing something to the training data so that these models either malfunction or they function in such a way that satisfies your malicious intent.

"The team that I work with has done it at ORNL. It has been done across the research community. We recently had a paper published when we were able to poison large language models to confuse on classifying certain pieces of text."



Evasion attacks

Unfortunately, you can often accomplish the same end by altering the image rather than poisoning the model. This is called an evasion attack.

If you change an image of yourself just so, you can fool an AI into thinking the image is of someone else while a person thinks it still looks like you. Such attacks have been amply demonstrated by researchers at ORNL and elsewhere.

Begoli said he and his colleagues demonstrated this technique on a trip abroad.

"On our recent visit to Alan Turing Institute, we were able to walk into the United Kingdom, and no human ever checked our passports — from Charlotte [North Carolina] to Heathrow [in London]. It was all done via biometrics. You walk in, the biometric scanner scans your face, it scans your passport, and if all matches, it lets you through.

"In the center, and in support of some our security programs, we were able to modify the photo used in these scenarios in such a way that the photo looked like me, but it had hidden features on the photo that confused the biometric system into thinking it was somebody else."

Unfortunately, he said, this type of attack is already widely in use.

"This is already done by human traffickers," he said, "but you can also imagine a terrorist who's on the terrorist watchlist and wants to be able to pass through safety checks or security checks, border controls."

Deepfakes

Just as you can create made-to-order AI images, you can do the same with videos. Make that fake image or video depict a real person, and you have a deepfake.

Deepfakes have been around for a while, and the technology does have harmless applications, but deepfakes can also be used for the darkest of purposes, from compromising videos used in blackmail schemes to fake news aimed at destabilizing political power.

"From a public safety concern, it's probably the most concerning area," Begoli said, "because it has immediate impact. Somebody can take a picture of somebody else and present it in all kinds of fabricated, deeply compromising situations."

Deepfakes are getting more and more convincing, but the most shocking aspect may be how cheap and easy they are to create.

"We demonstrated that one can generate hyperrealistic deepfakes for about $20," Begoli said. "It takes two hours, and you can generate a deepfake that looks like me and sounds like me. This was done by my [ORNL] colleague Sudarshan Srinivasan, who took a YouTube video of my talk and altered it in such a way that we created some deeply convincing, digital, proverbial Frankenstein that shared both of our vocal and facial features."

Begoli noted that there are techniques for detecting deepfakes, some being developed at ORNL, but such analysis is typically too little, too late.

"If you think about the target population — elderly, less educated in technology, people from across the world — it has major implications, political implications, national security implications. Somebody can have a false campaign and generate text and video and speech and all kinds of things for the people who are not as familiar."

To date, he said, the most effective tool against deepfakes may be public education.

"We're engaged a lot in educating the public and government on what deepfakes are, how concerning they are," he said. "What are the techniques to either detect and prevent, or in some instances it's honestly damage control, to be ready to mitigate the damage that may come out of having one spread around."

Scientific computing must be trustworthy

Scientists don't necessarily have the same attitude toward artificial intelligence as the rest of us. A skeptical crowd, they are not content to be given an answer. Instead, they need to know how the answer was arrived at and how confident they should be with it.

These questions get at trustworthiness in artificial intelligence, and they're not ones that AI models so far are very good at answering.

"AI models can give you results, but we want to associate uncertainties with outputs," Balaprakash said. "It's not like one model will be giving trustworthy results for all the cases. There are certain cases when it will not be trustworthy, and it's important to identify when it is not trustworthy and communicate that to the end user."

The second point of trustworthiness has to do with causal reasoning. Researchers need to know how a model came to its conclusion. Gina Tourassi, ORNL's associate laboratory director for computing and computational sciences, said the issue lies in an AI problem-solving technique known as heuristics.

"Trustworthiness is all about developing methodologies to help us understand the reasoning of these models in terms of how they get to an answer. … But I wouldn't say that I have seen tremendous progress in this particular field. It seems that the community puts a lot more emphasis on pushing the performance of the models by innovating, but innovating in terms of those heuristics, not really trying to understand why the model got better."

In fact, AI models sometimes provide answers that are false or misleading. **Generally known as hallucinations**, these responses can range from hilarious to dangerous.

Case in point: Here is a ChatGPT exchange from Sept. 30, 2024, in which the AI model was asked for a relevant quote from Tennessee native and technology supporter Al Gore.

The prompt:

Share some quotes from Al Gore on the promise of the internet. Include citations.

ChatGPT's response, in part:

"The Internet is a great tool for communicating, sharing ideas, and creating a community, but it can also be a tool for educating our children and empowering our citizens. It is the great equalizer, offering everyone access to the same information."
— (Speech at the Superhighway Summit, 1994)

This quote eloquently encapsulated early excitement and optimism over a powerful new technology. The problem was, it couldn't be verified. So after pasting the quote back into ChatGPT and asking when Gore had said it, this was the response (edited for conciseness):

It seems that the specific quote you mentioned ... does not correspond to a verifiable, widely cited speech or statement directly attributed to Al Gore with that exact wording.

However, the sentiment expressed in the quote closely aligns with themes Gore discussed during the 1990s, particularly in his advocacy for the internet as part of the "information superhighway."

In other words, it was not a quote from Al Gore, but it could have been a quote from Al Gore.

For Shaun Gleason, director of ORNL's Partnerships Office, though, such answers — while misleading — have the potential to lead researchers in promising directions, from a scientific viewpoint, at least.

"The AI system is just generating content based on what you fed it," he said. "It's not actually hallucinating. But you could think of a hallucination as a hypothesis that no one's really thought of or tested, maybe. It could be a new idea that makes no sense or makes perfect sense, and that's important for scientific discoveries."

AI systems must be energy efficient

Artificial intelligence is unbelievably powerful. It is also unbelievably power-hungry. According to the **World Economic Forum**, AI's power usage is growing between 26% and 36% a year. At that rate, AI in 2028 will use as much power as Iceland used in 2021.

The size of an AI model is typically conveyed by the number of its parameters — essentially the connections within the model. In general, the more the parameters, the more powerful and accurate an AI model will be.

Current models range from a few parameters up to trillions. ChatGPT 3, released in November 2022, had 175 billion parameters. Presumably, later versions have more, but its maker, OpenAI, isn't saying. Google's Gemini 1.5 Ultra has 1.56 trillion parameters.

It takes a tremendous amount of energy to train these large models. According to Gleason, it would take three to four months to train a trillion-parameter model on the lab's Frontier supercomputer, using the entire machine and running constantly day and night.

Keep in mind that Frontier, while being among the world's most powerful supercomputers, is also one of its most energy-efficient, ranking No. 22 on the GREEN500 list. Yet it does **consume 21 megawatts of power**, roughly the same as the city of Oak Ridge, Tennessee.

"That's why energy-efficient AI is such an important topic," Gleason said. "How do we build energy-efficient systems? How do we develop AI systems that, from hardware, software, and workflow perspectives, are very energy efficient.

"And are there alternative competing platforms like neuromorphic chips or other alternative computational hardware platforms that are much more energy efficient that we could use to train and deploy AI systems?"

GPUs drove the AI revolution, but their days may be numbered because of their power consumption.

Among the replacements being weighed for GPU-heavy AI systems are neuromorphic chips, inspired by the human brain, and quantum computers. Gleason noted that ORNL has research groups working on quantum machine learning.

"There's a lot of research going on in quantum machine learning right now and quantum computing for AI," Gleason said. "One of the reasons is that quantum computers, if they can build them large enough with enough error correction, can be much faster than classical computers for some tasks such as training large AI models. But the other benefit is they're expected to be tremendously energy efficient relative to classical computers."

Transcending the GPU and finding the right machinery can go a long way toward making AI sustainable, but it's not the whole game.

Tourassi stresses that the solution will also involve new and improved AI algorithms and software.

"There is no magic bullet," she said. "The hardware technology roadmaps show promise, but we have a long way to go with some of that hardware. The AI Initiative, though, is focused on more than just hardware. We need to be thinking about the three pillars: the hardware, the software and the algorithms.

"In terms of software, how can we rewrite codes so that they can execute calculations in a more energy-efficient way? And the third component is algorithms, relying on math and different kinds of algorithms that can help us get where we need to go, again, with a certain power envelope."

Tourassi also emphasized that AI's energy challenge goes beyond the energy required to train a large model, ravenous though those models are. We also need to look at the energy being used as the models interact with users. That problem will likely have to be approached in different ways.

"You can imagine every day millions of people asking billions of questions. Each one of those takes a little bit of power, but collectively these billions of requests on a day-to-day basis require a lot of energy. So how we tackle energy efficiency for a long training of a model, versus how we tackle energy efficiency for billions of short sprints, is a different challenge, and both are worthy of investigation. And that's what our laboratory is investing in," she said.

Looking to the future

When we look ahead and try to anticipate the direction artificial intelligence will take, several things become clear: AI is going to keep gaining steam, it's going to weave its way more deeply into every aspect of society, and it's going to come with serious challenges, from terrorists and blackmailers to power-hungry computers.

It has also become clear that AI and AI security research have turned into very promising fields of study.

"I think that over the next five to 10 years, this will probably become one of the most important fields within AI," Begoli said, "not just within research but within the commercial and federal sectors. My bigger concern is, will we have enough people to work in this space?"

Tourassi agreed that bringing in new AI experts will be key, certainly at ORNL and the other Department of Energy facilities.

"All of these developments and advances are not possible without a highly trained and competent workforce," she said. "There is tremendous demand for AI-skilled scientists and engineers across the complex."

When these new experts join ORNL, they will find a world-class organization working to ensure we don't repeat the mistakes of the past.

"Despite all these challenges, when developed and deployed responsibly, AI holds tremendous potential to bring about positive changes in our world," Balaprakash said. "We envision that within the next decade, we will witness major transformations across various fields of science and technology with secure, trustworthy, and energy-efficient AI. No other technology offers such immense potential, and it is worth striving for."

Continue reading ORNL Review: Turning AI into something we can trust

UT-Battelle manages ORNL for the Department of Energy's Office of Science, the single largest supporter of basic research in the physical sciences in the United States. The Office of Science is working to address some of the most pressing challenges of our time. For more information, please visit **energy.gov/science**.

###