Testimony of Bill Dally, Chief Scientist and Head of Research, NVIDIA

Before the
U.S. Senate Judiciary Committee
Subcommittee on Privacy, Technology, and the Law

Hearing on "Oversight of A.I.: Legislating on Artificial Intelligence"
September 12, 2023

Chairman Blumenthal, Ranking Member Hawley, and esteemed Judiciary Committee members, thank you for the privilege to testify today. I am NVIDIA's Chief Scientist and Head of Research, and I am delighted to provide our perspective on our artificial intelligence (AI) journey and future.

Our History

NVIDIA is at the forefront of accelerated computing and generative AI, technologies that have the potential to transform industries, address global challenges, and profoundly benefit society.

Since our founding in 1993, we have been committed to developing technology to empower people and improve the quality of life worldwide.

- 1993: Founded to bring 3D graphics to the gaming and multimedia markets.
- 1999: Invented the GPU, the graphics processing unit, which would later reshape the computing industry.
- 2006: Introduced the CUDA API, providing a GPU programming capability to researchers for applications in every field, including science, research, and enterprise.
- 2012: Spurred AI research and supported the breakthrough AlexNet neural network.
- 2018: Reinvented computer graphics with NVIDIA RTX, the first GPU capable of real-time ray tracing. From 2012 to 2022, pioneered architecture innovations resulting in a 1000x increase in AI performance, enabling modern generative AI.

Today, over 40,000 companies use NVIDIA platforms – across media and entertainment, scientific computing, healthcare, financial services, internet services, automotive, and manufacturing – to solve the world's most difficult challenges and bring new products and services to consumers worldwide.

We provide over three hundred libraries of accelerated code built on CUDA, our parallel computing developer platform and application programming interface. These libraries help developers take advantage of GPU-accelerated platforms. Today, our GPUs accelerate thousands of applications in every field, benefitting enterprises, consumers, and academics worldwide. Our ecosystem has attracted a global community of over 4 million developers, making NVIDIA a globally adopted computing platform.

Our Long AI Journey

AI has captured headlines in recent months, but AI did not arrive overnight.

Modern AI algorithms, based on deep neural networks, which are algorithms that use interconnected nodes in a layered structure, have been around since the 1950s. At that time, neural networks were largely academic pursuits; industry and policymakers paid them little attention. In the 80s, AI was dominated by logic programming and expert systems. A large industrial effort was made to commercialize expert systems, but they had little long-term impact.

At our founding in 1993, we were a 3D graphics startup, one of dozens of startups competing to create an entirely new market for accelerators to enhance graphics for computer games. At first, nobody seemed to see a need for another processor, much less one made by an untested startup. While others fell by the wayside, we forged ahead, inventing the GPU in 1999.

In contrast to CPUs, which are serial processors, our GPUs perform a massive number of calculations in parallel. When we launched the GPU for high-end gaming graphics, we recognized that GPUs could theoretically accelerate any application that could benefit from massively parallel processing.

But we needed to build an ecosystem and would have to convince developers to write applications for our products, one developer at a time. We firmly believed in our vision and set to work on it. In 2006, we unveiled the CUDA programming model, opening the parallel processing capabilities of GPUs to science and research. Even with CUDA, success did not come overnight. One by one, we worked with researchers and developers to teach the benefits of GPUs and CUDA.

We kept our nose to the grindstone, investing everything we could in research and development and building the ecosystem, even in lean years. We continued to refine CUDA and advance GPU technology, convinced that if we could innovate at the speed of light, the applications and the benefits to society would follow.

Over time, developers and researchers saw that accelerated computing could save time, energy, and money in various fields. As researchers learned to use CUDA, they realized that GPUs could accelerate the neural network algorithms that had previously resided only in textbooks. Neural networks require massively parallel processing, the raison d'etre of the GPU.

Thirteen years after NVIDIA invented the GPU, in 2012, researchers trained the AlexNet computer vision model with 14 million images on a pair of NVIDIA GPUs. The AlexNet model won the ImageNet challenge— an annual contest for image recognition algorithm accuracy — by a wide margin, igniting a flurry of research, largely performed on our GPU-powered platforms.

AlexNet didn't result in an immediate wave of commercial AI applications, but it left no doubt that we were on the right track. We redoubled our investment in our GPU-accelerated platforms and systems for the world's leading AI researchers. And each year, researchers clamored for more.

Five years after AlexNet, researchers using NVIDIA platforms invented the transformer model. After another five years, OpenAI scaled this model and its data set to release the GPT-3 large language model running on thousands of NVIDIA GPUs. This achievement led to the creation of ChatGPT, delighting and amazing users worldwide. Today, researchers across the globe innovate on NVIDIA GPUs.

AI's Potential and Responsible Design

As we look toward the future, we are thrilled to see generative AI's enormous potential, thanks to the remarkable developers who had faith in NVIDIA platforms. We are grateful to them and proud that our collective efforts have led to advances in AI that will revolutionize healthcare, medical research, education, business, and beyond.

For example, healthcare collaborations with academic institutions have led to large language models that can accurately predict a patient's risk of 30-day readmission, among other clinical outcomes. Imagine a world where AI-driven chatbots assist healthcare providers, helping them ask the right follow-up questions and ensuring that every possible issue is considered and evaluated and that the treatment is the best possible, meets all patient needs, and is explained to the patient in a way they can understand. AI will help radiologists correctly and rapidly interpret scans, dramatically reducing false results and saving thousands of lives. Medical AI will not eliminate the need for, or value of, human judgment, expertise, and care. Still, it will augment and raise the baseline abilities of human practitioners and provide a safety net to protect patients and practitioners alike. AI will never replace your human doctor or nurse, nor should it. But it will dramatically help them provide more personalized, accurate, efficient, and accessible healthcare.

In education, generative AI will expand access and uplift underserved communities worldwide. Teachers at all levels will be able to consult generative AI as a teaching assistant on nearly every topic, empowering educators to reach all students. Generative AI will be a valuable assistant for teachers in grading tests, creating new problem sets, and tutoring students who need more attention. As with healthcare, AI will never eliminate the need for students to connect with teachers, but it will enhance the experience for educators, students, and parents worldwide.

Generative AI can empower new tools to help businesses operate more efficiently and ensure compliance with financial, employment, commercial, and other laws. Just as an AI application can assist a doctor or an educator, an AI application can serve as a valuable assistant to ensure that the books and records of a company are in good order and reported accurately, to identify and mitigate fraud, and to ensure equal access and fair treatment in employment. In other words, we can use AI as a real-time auditor to help us identify and remedy fraud and unwanted bias *before* it causes harm. Again, AI will not eliminate the need for human professionals in business, financial, legal, and employment fields—rather it will empower them to do their jobs better than previously imaginable.

In our own field of computing, generative AI will make programmers and designers more productive, helping to reduce a workforce shortage that is challenging our country's ability to compete in this critical industry. Furthermore, AI-assisted design tools will be developed to provide real-time assistance to software developers—highlighting potential vulnerabilities before they are deployed, and in near real-time, suggesting secure alternatives.

AI will also be critical to the U.S.'s effort to cut greenhouse gases and combat climate change. AI will make every industry more efficient, and that efficiency will drive a reduction in greenhouse gas emissions. For example, AI will enhance and add intelligence to factories, transportation, logistics, HVAC systems, power generation, power grids, construction, and agriculture.

AI-driven innovations and management will create other environmental benefits, including enhancing water quality and reducing pollution. For example, AI-enabled water quality monitoring will swiftly detect and drive mitigation solutions to hazardous conditions and spills in our waterways and reservoirs. AI-enhanced forest monitoring will detect forest pest infestations, allowing for prompt treatment and mitigation. AI-powered air quality monitoring will detect and help mitigate toxic leaks. AI will enable targeted application of herbicides, dramatically reducing the chemical burden in our fields.

And, of course, AI applied to common computational use cases will dramatically reduce global greenhouse gas emissions while simultaneously adding significantly to the gross domestic product of nations that adopt it. Although significant energy is required to train models, once trained, running AI models saves energy. For example, AI-based simulations for fluid dynamics, weather, and climate are up to 100,000 times more efficient than conventional numerical simulations, allowing the same computation to be run at a tiny fraction of the energy use.

We recognize that, like any new product or service, AI products and services have risks. As it should be, the uses of AI systems are subject to our nation's laws and regulations, and everyone who makes, uses, or sells AI-enabled products and services is responsible for their conduct.

For example, like any other medical tool, the use or deployment of a medical AI chatbot must comply with the laws and regulations overseen by the FDA. Likewise, employment decisions must comply with all laws regarding bias and discrimination, whether assisted by an AI application or not. In financial services, lending decisions are subject to strict regulation—no exceptions for AI tools exist. Every endeavor that poses a serious risk of harm— from driving, flying, medicine, banking, engineering, and even legal services— is subject to licensing requirements that protect the public and promote quality, trustworthy products and services. AI-enabled services in high-risk sectors should be subject to licensing requirements as well. Other applications, with less risk of harm, may need less stringent licensing and/or regulation.

Benefitting from clear, stable, and thoughtful regulation, AI developers will work to benefit society while making our products and services as safe as possible. As we deploy AI more broadly, we can and will continue to identify and address risks— whether data privacy issues, undesired algorithmic biases, or potentials for misuse.
Efforts like NeMo Guardrails, NVIDIA's newly released open-source software, showcase our dedication to responsible AI advancement. NeMo Guardrails empowers developers to guide generative AI applications in producing accurate, appropriate, and secure text responses. Additionally, NVIDIA has established an internal model risk management guidance, ensuring a comprehensive assessment and management of risks associated with AI models.

Navigating the AI Frontier

No discussion of AI would be complete without addressing what are often described as "frontier AI models."

When we refer to frontier AI models, we are describing the next-generation, gigantic-scale models that vastly exceed the capabilities of anything we have today. Frontier AI models are not models like current implementations of ChatGPT, Bard, Ernie, or Llama, which have been exhaustively explored by developers and millions of users, and which have well-known and understood capabilities and limitations.

Depending on their size and training strategies the developers use, frontier AI models may possess unexpected, difficult-to-detect new capabilities. As a result, we must be especially thoughtful in our approach to developing and deploying frontier AI models, ensuring that we do not unleash models before they are safe, accurate, reliable, and doing exactly what we want them to do.

Some pundits have expressed fear that frontier AI models will evolve into uncontrollable "artificial general intelligence" products that will escape our control and cause harm. Fortunately, uncontrollable artificial general intelligence is science fiction, not reality. While we may not be able to predict everything that a given AI model *will* do, the way models are built and connected limits what they *can* do.

At its core, AI is a software program, not a nuclear reactor. AI is limited by its training, the inputs provided to it, and the nature of its output. When we create an AI, we first decide what tasks we want our AI to perform, and we train it to perform those tasks.

For example, suppose we train an AI to answer questions – type in a query, and the AI will provide an answer. That AI is not trained to or connected in a way that enables it to drive cars, fly airplanes, or control the power grid. And even if the AI could answer questions about how to do those tasks, the AI would have no ability to commandeer a car or an aircraft. The AI resides exactly where we put it, can do only what we train it to do, and can only affect what its outputs are connected to.

As a result, we humans will always decide how much decision-making power to cede to AI models. The AI models will never seize power by themselves. They will have only the power we give them. We can keep a human in the loop for any activity we wish.

We have no doubt that industry and government can work collaboratively to ensure that AI models are developed and deployed ethically and responsibly. The Administration's voluntary commitments, endorsed by NVIDIA and many of the companies engaged in frontier AI model development, are a solid start.

Promoting Innovation and Democratizing AI

While a few companies tend to dominate the AI headlines, we should recognize that the next great AI applications may come from anywhere in the world.

Today, thousands of startups worldwide are working diligently on foundational models and AI techniques, striving to create new tools and software that will benefit society in a myriad of ways. They work on GPU-powered systems deployed in workstations and the cloud. That work should be celebrated and encouraged and is well within the reach of today's regulatory and legal framework. These startups can and should be regulated as any other business.

So long as we are thoughtful and measured, we can ensure the safe, trustworthy, and ethical deployment of AI systems without suppressing innovation by researchers, academics, and enterprises working on new applications today. And we can spur innovation by ensuring that AI tools are widely available to everyone, not concentrated in the hands of a few powerful firms.

We encourage policymakers to work with industry to address concerns regarding frontier AI models while spurring startups, researchers, and enterprises to innovate to benefit society as rapidly as possible – which leads to a final observation.

Safe and trustworthy AI will require multilateral cooperation on norms and regulations, with the participation of every major power, or it will not be effective. This observation follows from two foundational truths.

First, the AI genie is already out of the bottle. The technology industry is international in scope, with many competitors worldwide. AI algorithms are widely published and available to all. AI software can be transmitted anywhere in the world at the press of a button, and many AI development tools, frameworks, and foundational models are open-sourced, spurring companies in every country to compete to create new AI-enabled tools and services.

Second, no nation, and certainly no company, controls a "chokepoint" to AI development. Leading U.S. computing platforms are competing with companies from around the world. While U.S. companies may currently be the most energy-efficient, cost-efficient, and/or easiest to use, they are not the only viable alternatives for developers abroad. Other nations are developing AI systems, with or without U.S. components, and they will offer those applications in the worldwide market. To ensure safe and trustworthy AI, we must engage with governments and stakeholders around the world on norms and regulations to ensure responsible and trustworthy AI.

The United States is in a remarkable position today, and with your help, we can lead well into the future.

Thank you for the opportunity to testify before this Committee. NVIDIA stands ready to work with you to ensure that the development and deployment of generative AI and accelerated computing serve the best interests of all.