2023 Predictions: AI That Bends Reality, Unwinds the Golden Screw and Self-Replicates

15 NVIDIA AI experts predict digital twins and generative AI are set to advance enterprise goals and consumer needs even as the world enters a third year of planning uncertainty.

Author: Cliff Edwards

After three years of uncertainty caused by the pandemic and its post-lockdown hangover, enterprises in 2023 — even with recession looming and uncertainty abounding — face the same imperatives as before: lead, innovate and problem solve.

AI is becoming the common thread in accomplishing these goals. On average, 54% of enterprise AI projects made it from pilot to production, according to a recent Gartner survey of nearly 700 enterprises in the U.S., U.K. and Germany. A whopping 80% of executives in the survey said automation can be applied to any business decision, and that they're shifting away from tactical to strategic uses of AI.

The mantra for 2023? Do more with less. Some of NVIDIA's experts in AI predict businesses will prioritize scaling their AI projects amid layoffs and skilled worker shortages by using cloud-based integrated software and hardware offerings that can be purchased and customized to any enterprise, application or budget.

Cost-effective AI development also is a recurring theme among our expert predictions for 2023. With Moore's law running up against the laws of physics, installing on-premises compute power is getting more expensive and less energy efficient. And the Golden Screw search for critical components is speeding the shift to the cloud for developing AI applications as well as for finding data-driven solutions to supply chain issues.

Here's what our experts have to say about the year ahead in AI:

ANIMA ANANDKUMAR Director of ML Research, and Bren Professor at Caltech

Digital Twins Get Physical: We will see large-scale digital twins of physical processes that are complex and multi-scale, such as weather and climate models, seismic phenomena and material properties. This will accelerate current scientific simulations as much as a million-x, and enable new scientific insights and discoveries.

Generalist AI Agents: AI agents will solve open-ended tasks with natural language instructions and large-scale reinforcement learning, while harnessing foundation models — those large AI models trained on a vast quantity of unlabeled data at scale — to enable agents that can parse any type of request and adapt to new types of questions over time.

MANUVIR DAS Vice President, Enterprise Computing

Software Advances End AI Silos: Enterprises have long had to choose between cloud computing and hybrid architectures for AI research and development — a practice that can stifle developer productivity and slow innovation. In 2023, software will enable businesses to unify AI pipelines across all infrastructure types and deliver a single, connected experience for AI practitioners. This will allow enterprises to balance costs against strategic objectives, regardless of project size or complexity, and provide access to virtually unlimited capacity for flexible development.

Generative AI Transforms Enterprise Applications: The hype about generative AI becomes reality in 2023. That's because the foundations for true generative AI are finally in place, with software that can transform large language models and recommender systems into production applications that go beyond images to intelligently answer questions, create content and even spark discoveries. This new creative era will fuel massive advances in personalized customer service, drive new business models and pave the way for breakthroughs in healthcare.

KIMBERLY POWELL Vice President, Healthcare

Biology Becomes Information Science: Breakthroughs in large language models and the fortunate ability to describe biology in a sequence of characters are giving researchers the ability to train a new class of AI models for chemistry and biology. The capabilities of these new AI models give drug discovery teams the ability to generate, represent and predict the properties and interactions of molecules and proteins — all in silicon. This will accelerate our ability to explore the essentially infinite space of potential therapies.

Surgery 4.0 Is Here: Flight simulators serve to train pilots and research new aircraft control. The same is now true for surgeons and robotic surgery device makers. Digital twins that can simulate at every scale, from the operating room environment to the medical robot and patient anatomy, are breaking new ground in personalized surgical rehearsals and designing AI-driven human and machine interactions. Long residencies won’t be the only way to produce an experienced surgeon. Many will become expert operators when they perform their first robot-assisted surgery on a real patient.

DANNY SHAPIRO Vice President, Automotive

Training Autonomous Vehicles in the Metaverse: The more than 250 auto and truck makers, startups, transportation and mobility-as-a-service providers developing autonomous vehicles are tackling one of the most complex AI challenges of our time. It’s simply not possible to encounter every scenario they must be able to handle by testing on the road, so much of the industry in 2023 will turn to the virtual world to help.

On-road data collection will be supplemented by virtual fleets that generate data for training and testing new features before deployment. High-fidelity simulation will run autonomous vehicles through a virtually infinite range of scenarios and environments. We’ll also see the continued deployment of digital twins for vehicle production to improve manufacturing efficiencies, streamline operations and improve worker ergonomics and safety.

Moving to the Cloud: 2023 will bring more software-as-a-service (SaaS) and infrastructure-as-a-service offerings to the transportation industry. Developers will be able to access a comprehensive suite of cloud services to design, deploy and experience metaverse applications anywhere. Teams will design and collaborate on 3D workflows — such as AV development simulation, in-vehicle experiences, cloud gaming and even car configurators delivered via the web or in showrooms.

Your In-Vehicle Concierge: Advances in conversational AI, natural language processing, gesture detection and avatar animation are making their way to next-generation vehicles in the form of digital assistants. This AI concierge can make reservations, access vehicle controls and provide alerts using natural language understanding. Using interior cameras, deep neural networks and multimodal interaction, vehicles will be able to ensure that driver attention is on the road and ensure no passenger or pet is left behind when the journey is complete.

REV LEBAREDIAN Vice President, Omniverse and Simulation Technology

The Metaverse Universal Translator: Just as HTML is the standard language of the 2D web, Universal Scene Description is set to become the most powerful, extensible, open language for the 3D web. As the 3D standard for describing virtual worlds in the metaverse, USD will allow enterprises and even consumers to move between different 3D worlds using various tools, viewers and browsers in the most seamless and consistent fashion.

Bending Reality With Digital Twins: A new class of true-to-reality digital twins of goods, services and locations is set to offer greater windfalls than their real-world counterparts. Imagine selling many virtual pairs of sneakers in partnership with a gaming company that are simply undergoing design testing — long before sending the pattern to manufacturing. Companies also stand to benefit by saving on waste, increasing operational efficiencies and boosting accuracy.

RONNIE VASISHTA Senior Vice President, Telecoms

Cutting the Cord on AR/VR Over 5G Networks: While many businesses will move to the cloud for hardware and software development, edge design and collaboration also will grow as 5G networks become more fully deployed around the world. Automotive designers, for instance, can don augmented reality headsets and stream the same content they see over wireless networks to colleagues around the world, speeding collaborative changes and developing innovative solutions at record speeds. 5G also will lead to accelerated deployments of connected robots across industries — used for restocking store shelves, cleaning floors, delivering pizzas and picking and packing goods in factories.

RAN in the Cloud: Network operators around the world are rolling out software-defined virtual radio access network 5G to save time and money as they seek faster returns on their multibillion-dollar investments. Now, they're shifting away from bespoke L1 accelerators to 100% software-defined and full-stack, 5G-baseband acceleration that includes L2, RIC, Beamforming and FH offerings. This shift will lead to an increase in the utilization of RAN systems by enabling multi-tenancy between RAN and AI workloads.

BOB PETTE Vice President, Professional Visualization

An Industrial Revolution via Simulation: Everything built in the physical world will first be simulated in a virtual world that obeys the laws of physics. These digital twins — including large-scale environments such as factories, cities and even the entire planet — and the industrial metaverse are set to become critical components of digital transformation initiatives. Examples already abound: Siemens is taking industrial automation to a new level. BMW is simulating entire factory floors to optimally plan manufacturing processes. Lockheed Martin is simulating the behavior of forest fires to anticipate where and when to deploy resources. DNEG, SONY Pictures, WPP and others are boosting productivity through globally distributed art departments that enable creators, artists and designers to iterate on scenes virtually in real time.

Rethinking of Enterprise IT Architecture : Just as many businesses scrambled to adapt their culture and technologies to meet the challenges of hybrid work, the new year will bring a re-architecting of many companies' entire IT infrastructure. Companies will seek powerful client devices capable of tackling the ever-increasing demands of applications and complex datasets. And they'll embrace flexibility, moving to burst to the cloud for exponential scaling. The adoption of distributed computing software platforms will enable a globally dispersed workforce to collaborate and stay productive under the most disparate working environments.

Similarly, complex AI model development and training will require powerful compute infrastructure in the data center and the desktop. Businesses will look at curated AI software stacks for different industrial use cases to make it easy for them to bring AI into their workflows and deliver higher quality products and services to customers faster.

AZITA MARTIN Vice President, AI for Retail, Consumer Packaged Group and Quick-Service Restaurants

Tackling Shrinkage: Brick-and-mortar retailers perennially struggle with a commonplace problem: shrinkage, the industry parlance for theft. As more and more adopt AI-based services for contactless checkout, they'll seek sophisticated software that combines computer vision with store analytics data to make sure what a shopper rings up is actually the item being purchased. The adoption of smart self-tracking technology will aid in the development of fully automated store experiences and help solve for labor shortages and lost income.

AI to Optimize Supply Chains: Even the most sophisticated retailers and e-commerce companies had trouble the past two years balancing supply with demand. Consumers embraced home shopping during the pandemic and then flocked back into brick-and-mortar stores after lockdowns were lifted. After inflation hit, they changed their buying habits once again, giving supply chain managers fits. AI will enable more frequent and more accurate forecasting, ensuring the right product is at the right store at the right time. Also, retailers will embrace route optimization software and simulation technology to provide a more holistic view of opportunities and pitfalls.

MALCOLM DEMAYO Vice President, Financial Services

Better Risk Management: Firms will look for opportunities like accelerated compute to drive efficiencies. The simulation techniques used to value risk in derivatives trading are computationally intensive and typically consume large swaths of data center space, power and cooling. What runs all night on traditional compute will run over a lunch break or faster on accelerated compute. A real-time value of sensitivities will enable firms to better manage risk and improve the value they deliver to their investors.

Cloud-First for Financial Services: Banks have a new imperative: get agile fast. Facing increasing competition from non-traditional financial institutions, changing customer expectations rising from their experiences in other industries and saddled with legacy infrastructure, banks and other institutions will embrace a cloud-first AI approach. But as a highly regulated industry that requires operational resiliency, an industry term that means your systems can absorb and survive shocks (like a pandemic), banks will look for open, portable, hardened, hybrid solutions. As a result, banks are obligated to purchase support agreements when available.

CHARLIE BOYLE Vice President, DGX systems

AI Becomes Cost-Effective With Energy-Efficient Computing: In 2023, inefficient, x86-based legacy computing architectures that can’t support parallel processing will give way to accelerated computing solutions that deliver the computational performance, scale and efficiency needed to build language models, recommenders and more.

Amidst economic headwinds, enterprises will seek out AI solutions that can deliver on objectives, while streamlining IT costs and boosting efficiency. New platforms that use software to integrate workflows across infrastructure will deliver computing performance breakthroughs — with lower total cost of ownership, reduced carbon footprint and faster return on investment on transformative AI projects — displacing more wasteful, older architectures.

DAVID REBER Chief Security Officer

Data Scientists Are Your New Cyber Asset: Traditional cyber professionals can no longer effectively defend against the most sophisticated threats because the speed and complexity of attacks and defense have effectively exceeded human capacities. Data scientists and other human analysts will use AI to look at all of the data objectively and discover threats. Breaches are going to happen, so data science techniques using AI and humans will help find the needle in the haystack and respond quickly.

AI Cybersecurity Gets Customized: Just like recommender systems serve every consumer on the planet, AI cybersecurity systems will accommodate every business. Tailored solutions will become the No. 1 need for enterprises’ security operations centers as identity-based attacks increase. Cybersecurity is everyone’s problem, so we’ll see more transparency and sharing of various types of cybersecurity architectures. Democratizing AI enables everyone to contribute to the solution. As a result, the collective defense of the ecosystem will move faster to counter threats.

KARI BRISKI Vice President, AI and HPC Software

The Rise of LLM Applications: Research on large language models will lead to new types of practical applications that can transform languages, text and even images into useful insights that can be used across a multitude of diverse organizations by everyone from business executives to fine artists. We'll also see rapid growth in demand for the ability to customize models so that LLM expertise spreads to languages and dialects far beyond English, as well as across business domains, from generating catalog descriptions to summarizing medical notes.

Unlabeled Data Finds Its Purpose : Large language models and structured data will also extend to the reams of photos, audio recordings, tweets and more to find hidden patterns and clues to support healthcare breakthroughs, advancements in science, better customer engagements and even major advances in self-driving transportation. In 2023, adding all this unstructured data to the mix will help develop neural networks that can, for instance, generate synthetic profiles to mimic the health records they've learned from. This type of unsupervised machine learning is set to become as important as supervised machine learning.

The New Call Center: Keep an eye on the call center in 2023, where adoption of more and more easily implemented speech AI workflows will provide business flexibility at every step of the customer interaction pipeline — from modifying model architectures to fine-tuning models on proprietary data and customizing pipelines. As the accessibility of speech AI workflows broadens, we'll see a widening of enterprise adoption and giant increase in call center productivity by speeding time to resolution. AI will help agents pull the right information out of a massive knowledge base at the right time, minimizing wait times for customers.

KEVIN DEIERLING Senior Vice President, Networking

Moore's Law on Life Support: As CPU design runs up against the laws of physics and struggles to keep up with Moore's law — the postulation that roughly every two years the number of transistors on microchips would double and create faster, more efficient processing — enterprises increasingly will turn to accelerated computing. They'll use custom combinations of CPUs, GPUs, DPUs and more in scalable data centers to innovate faster while becoming more cloud oriented and energy efficient.

The Network as the New Computing Platform: Just as personal computers combined software, hardware and storage into productivity-generating tools for everyone, the cloud is fast becoming the new computing tool for AI and the network is what enables the cloud. Enterprises will use third-party software, or bring their own, to develop AI applications and services that run both on-prem and in the cloud. They'll use cloud services operators to purchase the capacity they need when they need it, working across CPUs, GPUs, DPUs and intelligent switches to optimize compute, storage and the network for their different workloads. What's more, with zero-trust security being rapidly adopted by cloud service providers, the cloud will deliver computing as secure as on-prem solutions.

DEEPU TALLA Vice President, Embedded and Edge Computing

Robots Get a Million Lives: More robots will be trained in virtual worlds as photorealistic rendering and accurate physics modeling combine with the ability to simulate in parallel millions of instances of a robot on GPUs in the cloud. Generative AI techniques will make it easier to create highly realistic 3D simulation scenarios and further accelerate the adoption of simulation and synthetic data for developing more capable robots.

Expanding the Horizon: Most robots operate in constrained environments where there is limited to no human activity. Advancements in edge computing and AI will enable robots to have multi-modal perception for better semantic understanding of their environment. This will drive increased adoption of robots operating in brownfield facilities and public spaces such as retail stores, hospitals and hotels.

MARC SPIELER Senior Director, Energy

AI-Powered Energy Grid: As the grid becomes more complex due to the unprecedented rate of distributed energy resources being added, electric utility companies will require edge AI to improve operational efficiency, enhance functional safety, increase accuracy of load and demand forecasting, and accelerate the connection time of renewable energy, like solar and wind. AI at the edge will increase grid resiliency, while reducing energy waste and cost.

More Accurate Extreme Weather Forecasting: A combination of AI and physics can help better predict the world's atmosphere using a technique called Fourier Neural Operator. The FourCastNet system is able to predict a precise path of a hurricane and can also make weather predictions in advance and provide real-time updates as climate conditions change. Using this information will allow energy companies to better plan for renewable energy expenditures, predict generation capacity and prepare for severe weather events.

Original URL: https://blogs.nvidia.com/blog/2022/12/13/2023-ai-predictions/