Goldman Sachs Global Institute

When AI Learns How the World Works

APRIL 2026

George Lee
Co-Head, Goldman Sachs Global Institute

Dan Keyserling
Managing Director, Goldman Sachs Global Institute

How AI World Models Create an Operating System for Decision-Making

After a decade defined by systems that recognize patterns and predict text, the frontier of AI is shifting toward models that understand how the world works. The next advances in AI may come less from bigger models and more from systems that can simulate reality, test actions before taking them, and reason about consequences. This new category of models, known as world models, represents a quiet but decisive change in how machines become intelligent.

For the past few years, artificial intelligence has been defined by large language models (LLM). Trained on vast swathes of text, they learned to predict the next word with uncanny accuracy. From that simple objective emerged systems that write, translate, code, and converse with startling fluency. That achievement is real and transformative, but it also reveals a limitation to the current generation of AI models.

LLMs are powerful at completing patterns, but they lack the internal sense of the world those patterns describe. They respond well to prompts but struggle to reason through consequences or act reliably in environments where mistakes carry costs. This limitation has become clearer as these systems have been pushed beyond text. When they're asked to control robots, manage entire supply chains, or coordinate complex enterprise workflows, prediction alone proves insufficient. Intelligence, in these settings, requires more than correlation. It requires an internal model of how the world works.

Consider the layers in practice: An LLM can extract covenants from a stack of loan documents or draft an investment committee memo. A physical world model can simulate how a hurricane season reshapes insured-loss distributions across a reinsurance portfolio. A social world model can forecast how a policy shock cascades through markets and behavior. The most consequential decisions may eventually draw on all three capabilities—yet plenty of high-value financial tasks remain squarely in LLM territory today. What's changing is that building these natural evolutions of LLMs is no longer a fringe ambition. It has become a strategic priority for some of AI's most influential researchers.

Yann LeCun, who recently left his position as Chief AI Scientist at Meta, has made world models the centerpiece of his vision for artificial general intelligence and his new venture, AMI Labs. His Joint-Embedding Predictive Architecture (JEPA) framework explicitly aims to build machines that learn world models from observation, much as humans do, focusing on predicting abstract representations or concepts about what comes next without reconstructing exact details. Meanwhile, Fei-Fei Li, the Stanford professor whose ImageNet dataset helped catalyze the deep learning revolution, has founded a new venture focused on spatial intelligence. Her work at World Labs emphasizes that true intelligence requires not just recognizing objects in images but understanding how those objects exist in space, how they interact, and how they change over time.

In other words, instead of asking models to respond to inputs, researchers create internal representations of the world so they can run simulations inside them. These so-called world models allow systems to imagine outcomes before taking an action. They run mental experiments. They test possibilities. You could call it a primitive form of machine foresight. But the term *world model* hides an important distinction.

There are *two* kinds of worlds AI is learning to model. One is the physical world of gravity, friction, heat, and force. The other is a virtual or social world, populated by many interacting agents with goals, memories, and constraints. Each points toward a different frontier. Together, they hint at a deeper shift in what intelligence looks like.

> A world model is an internal simulator—it allows a system to ask a simple question repeatedly: If I do this, what happens next? Humans rely on this instinct constantly.

Put simply, a world model is an internal simulator—it allows a system to ask a simple question repeatedly: If I do this, what happens next? Humans rely on this instinct constantly. We picture a glass tipping before it falls. We imagine a meeting going badly before choosing our words. Until recently, machines could not do this well. Teaching a robot to recognize a cup is easy. Teaching it to pick one up without shattering it is hard. The real world is unforgiving. Objects have weight. Surfaces have friction. Liquids spill. Small errors compound quickly. For decades, robots worked best in carefully controlled environments, fenced off from human and real-world unpredictability. Even today's warehouse robots navigate mapped, partly rule-bound spaces. Physical world models promise something more: models that can handle the unstructured real world.

Instead of learning only from trial and error in the real world, physical world models enable machines to learn the rules that govern it. They absorb the logic of physics, thermodynamics, fluid dynamics, and material science. They practice inside simulations that mimic reality closely enough to matter.

These simulations aren't new. What is new is their scale and fidelity. Advances in computing, reinforcement learning, and synthetic data allow machines to run *millions* of imagined experiments before touching the real world. A robot can learn how to walk, grasp, or balance by failing thousands of times in a simulation, where failure is cheap. When it finally acts in the real world, it does so with a plan.

This approach has quietly unlocked progress in logistics, manufacturing, and autonomous systems. Warehouse robots navigate crowded spaces with fewer collisions (even in complete darkness). Machines adapt to unfamiliar objects instead of glitching. Autonomous vehicles rehearse edge cases long before they encounter them on the road. The critical advance is not better hardware (though that helps) but better internal models of reality.

If physical world models teach machines how the world behaves, *virtual* world models explore how people and institutions behave. Here, physics is social rather than mechanical. The forces are incentives, norms, information, and power.

These models consist of digital environments populated by many AI agents. Each agent has goals, memory, and the ability to reason. (Agents can even be assigned "personas" that mimic specific real-world behavioral profiles and characteristics.) They interact with one another over time. Out of those interactions emerge patterns. Some of those patterns are the result of random interaction, but some are the product of knowable features of the underlying systems.

What makes virtual world models especially powerful is their ability to approximate the behavior of real groups of people, not in aggregate but in interaction. Enterprises already spend enormous effort guessing how others will respond, how competitors will move, how markets will interpret signals, how boards will react under pressure. Today, those judgments rely on experience, static analysis, and intuition. Multi-agent simulations offer something closer to a living model of human systems. By populating digital environments with agents that reflect different incentives, constraints, and information sets, firms gain a higher-fidelity operating system for decision-making. Strategies can be tested against adaptive components. Governance structures can be stress-tested before crisis hits. In trading, corporate strategy, and board-level playing, the advantage lies less in faster answers and more in better rehearsal.

The practical applications are already visible. Many firms and governments use multi-person simulations to stress-test strategies before deploying them. Perhaps the most famous example is war games, in which military leaders use scenario-based simulation to explore how a military conflict might play out. Policymakers rehearse how information campaigns might spread through a population. Organizations rehearse crisis responses without bearing real-world cost. Imagine the very best planned simulation that relies on real humans to play certain parts—now imagine being able to replicate that same scenario in a digital environment *thousands* of times, observing all the variabilities in outcomes. Such is the power of world models.

These systems don't predict the future in any narrow sense; they're meant to reveal plausible futures and expose hidden dynamics. This distinction—between prediction and simulation—is important. Forecasting assumes a single correct outcome. World models reveal ranges, paths, and feedback loops. They show how systems behave under pressure and how individuals behave within those systems. For leaders, that is often more useful than a static estimate of an outcome.

At first glance, physical and virtual world models seem unrelated. One concerns robots and machines; the other concerns people and institutions. Yet they share a common logic. Both require AI to understand systems governed by constraints. Both emphasize causality over correlation. Both reward anticipation over reaction.

The parallels run deeper, and it is tempting to draw parallels between world models and how the human brain builds internal simulations of its environment. We do not have a single, monolithic system for understanding intelligence. Instead, the brain is a federation of specialized regions. The visual cortex processes images, Broca's area handles language production, the motor cortex coordinates movement, and the prefrontal cortex orchestrates planning and decision making. These systems evolved separately, yet they work in concert. Language informs motor planning, visual perception shapes verbal description, and abstract reasoning draws on embodied experience.

Just as the brain integrates specialized modules into coherent thought, advanced AI architectures will likely combine language models, physical simulators, and social reasoning engines into unified systems. Much of the current discourse pits LLMs against world models, asking which paradigm will prevail. The debate over which paradigm "wins" fundamentally misunderstands how complex intelligence emerges: not from a single dominant approach but from the orchestration of many. The most capable AI system of the future will likely integrate both, using language as an interface for instruction and explanation while relying on world models for planning and consequence-aware action.

Physical laws constrain motion. Social rules constrain behavior. Objects exert forces. Incentives do the same. In the latter two cases, intelligence emerges from understanding how local actions ripple outward. Seen this way, world models mark a transition of AI from pattern recognition to system understanding. This shift carries economic and strategic consequences.

Consider a future supply chain. Physical world models guide robots moving goods through warehouses and ports. Virtual world models simulate demand shocks, labor responses, and geopolitical disruptions. Decisions in one world inform actions in the other. Planning becomes continuous rather than episodic.

This is why world models matter now. The frontier of AI no longer lies only in larger models and more data. It lies in better representations of reality.

This raises a provocative question: Are we actually underestimating total AI spend? The demands and opportunities surrounding world models are not yet reflected in consensus supply-and-demand forecasts for AI infrastructure that are primarily focused on transformer-based LLMs. Current projections for compute, energy, and chip demand are largely built around the scaling of large language models, from data to training to inference-time compute. But if world models prove as important as laid out above—and operate in a manner that is complementary rather than substitutional to LLMs—could the implications be significant?

In the near term, world-model investment is likely to remain a fraction of total AI spend, especially given the current state and pace of LLMs in commercial adoption. But the trajectory matters: As these systems open and accelerate broad new horizons for simulation, robotics, autonomous systems, and strategic planning, the aggregate compute requirements could exceed what current forecasts anticipate.

In the near term, world-model investment is likely to remain a fraction of total AI spend, especially given the current state and pace of LLMs in commercial adoption. But the trajectory matters: As these systems open and accelerate broad new horizons for simulation, robotics, autonomous systems, and strategic planning, the aggregate compute requirements could exceed what current forecasts anticipate. That said, some of this complementary infrastructure overlaps. The same GPU clusters and inference platforms that serve LLMs can also train and run world models. But the similarities have limits: Simulation environments typically demand purpose-built data pipelines, synthetic data generators, and physics-based engines that go well beyond text corpora. The infrastructure story is one of partial overlap, not seamless reuse.

This might not come cheap. Critics note, correctly, that world models are computationally expensive. High-fidelity simulation, multi-agent interaction, and continuous planning consume far more compute than predicting the next word in a sentence. But cost alone is the wrong metric. In domains where mistakes are expensive and foresight creates leverage, the value of simulation compounds faster than its compute bill.

Building high-fidelity simulations of physical dynamics and complex social systems may sound prohibitively expensive, but it might be more feasible than intuition suggests. For example, the worlds modeled need not be complete, only relevant. A warehouse robot does not need to simulate weather patterns or geopolitics. A geopolitical crisis response simulation does not need to understand organic chemistry. World models can be scoped, compressed, and specialized. We can already observe this throughout the last year, which is why coding and mathematics tasks have seen such drastic improvements. The environment, specialized in coding and math, provides clear signals: A solution either compiles or passes tests, or it doesn't.

This technique, called reinforcement learning, teaches systems through reward and punishment and is thus already a necessary ingredient for improving the performance of today's LLMs. However, extending this concept to the physical and social domains—which world models seek to simulate—remains a complex and evolving area of research. The challenge lies in designing reward signals for environments where the criteria for success are less clearly defined and outcomes are delayed. No system is being systematically punished for dropping a glass, for misreading a war room, for choosing words that erode trust. When a model receives a reward for deliberating longer on a coding challenge and producing the correct result, verified instantly through automated tests, it learns to think. When it fails those tests, it's punished. This is immediate, precise, and scalable. The feedback loops that govern embodied and social intelligence remain largely absent from training regimes still primarily based on video data—encouraging systems that learn how scenes look rather than how they feel.

Forecasting exact compute needs is difficult, but the trajectory is clear. As foundation models gain efficiency and specialized accelerators proliferate, the cost of running sophisticated simulations continues to fall. The bottleneck is shifting from raw compute to the scope and quality of the simulation itself and how faithfully it captures dynamics that matter.

Several signals are worth watching. Investment is shifting from stand-alone models to full simulation environments. Synthetic data are beginning to outnumber real data in training regimes. Evaluation metrics are moving away from prediction and toward decision quality over time. Large organizations are building digital twins of operations, markets, and infrastructure.

Models trained on text seem to be able to exhibit a sort of understanding of our world. However, these models generate this understanding through second-order interpretation—they understand how our world works based on the data and text to which they have been exposed. They do not possess first-principles understanding of physics, motion, light, action/reaction, or other fundamental properties of our universe.

The implications of world models are subtle but profound. Intelligence, artificial or otherwise, is less about answers than about foresight. World models give machines something they have long lacked: a sense of consequence.

For much of its recent history, we've treated artificial intelligence as a system that produces answers. World models suggest something more ambitious.

For much of its recent history, we've treated artificial intelligence as a system that produces answers. World models suggest something more ambitious. They point toward machines that understand context, constraint, and consequence. If large language models give AI fluency, world models give it situational awareness. That shift alters the paradigm. Intelligence becomes less about generating plausible outputs and more about navigating structured realities. In that transition lies the next frontier—not bigger models but deeper ones, systems that reason inside worlds rather than merely describing them.

If this transition occurs, it will reshape the industry itself. Competitive advantage might depend as much on who trains the largest model as who builds the most faithful simulations of reality, physical, social, and economic. This is why world models are more than technical refinement. They signal a deeper change in what artificial intelligence is for and in how profoundly it may reorder the decisions that shape modern life. That change will demand new investment, new infrastructure, and new ways of measuring progress. Organizations that recognize this shift early will be better positioned not just to adopt AI but to deploy it where it matters most—in decisions that shape the real world.

Authors

GEORGE LEE

Co-Head
Goldman Sachs Global Institute

DAN KEYSERLING

Managing Director
Goldman Sachs Global Institute

Disclaimer