PRIMER

NATIONAL SECURITY AND POLICY IMPLICATIONS OF WORLD MODELS

LLMS WERE NEVER THE WHOLE OF AI, ONLY ITS FIRST COMMERCIAL FRONTIER

GABRIELLE TRAN

JULY 2026

IST Institute for SECURITY + TECHNOLOGY

National Security and Policy Implications of World Models:
LLMs were never the whole of AI, only its first commercial frontier

July 2026
Author: Gabrielle Tran
Design: Taylor White

About the Institute for Security and Technology

Uniting technology and policy leaders to create actionable solutions to emerging security challenges

Technology has the potential to unlock greater knowledge, enhance our collective capabilities, and create new opportunities for growth and innovation. However, insecure, negligent, or exploitative technological advancements can threaten global security and stability. Anticipating these issues and guiding the development of trustworthy technology is essential to preserve what we all value.

The Institute for Security and Technology (IST), the 501(c)(3) critical action think tank, stands at the forefront of this imperative, uniting policymakers, technology experts, and industry leaders to identify and translate discourse into impact. We take collaborative action to advance national security and global stability through technology built on trust, guiding businesses and governments with hands-on expertise, in-depth analysis, and a global network.

We work across three analytical pillars: the **Future of Digital Security**, examining the systemic security risks of societal dependence on digital technologies; **Geopolitics of Technology**, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; and **Innovation and Catastrophic Risk**, providing deep technical and analytical expertise on technology-derived existential threats to society.

Learn more: https://securityandtechnology.org/

Contents

MARY'S ROOM 1

Introduction 2

What is a World Model? 3

Today's **Generative** World Models 3

Today's **Non-Generative** World Models: JEPA-trained World Models 4

Use Cases and Comparative Advantages 6

Specific Implications: What world models mean for national security 8

Broader Implications: What could go wrong 9

Questions for Policymakers 11

Conclusion 13

MARY'S ROOM

In 1982, the philosopher Frank Jackson created a thought experiment he called "Mary's Room."

In this thought experiment, Mary is a gifted scientist who has lived her entire life in a black and white room. She can only investigate the world through the lens of a black and white television monitor. Despite her limits, she knows every (yes, every) fact about color, from wavelengths and how the retina processes light to the "expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue.'" One day, she leaves the room and sees color for the first time. Jackson asks: has Mary learned anything new?

According to Jackson, Mary has indeed learned something new. The thought experiment substantiated the view that conscious experience involves non-physical elements that cannot be captured by any amount of physical facts alone. Now, let's imagine Mary is an LLM.

Like Mary, that LLM has exhaustive propositional knowledge about color—in fact, the LLM possesses almost every fact ever documented about it. But in the same way that Mary learns something new and entirely irreducible to physical description when she steps outside and sees color, then an LLM would face the same wall she did. No matter how much text the LLM trains on, it can only ever consume the same knowledge that Mary already knew. Moreover, the LLM does not learn about color directly. It can only glean information from what people have written about color, which is not the same as what people know about color.

What does this thought experiment mean for scaling large language models? Some have argued that scaling LLMs alone will never produce the same kind of grounding that Mary gains when she opens the door. To achieve that grounding, LLMs would need to develop an internal representation of reality, one that can envision future states with an embedded sense of physics, 3D space, and cause and effect. This vision for internal representations of reality is the ambition behind a new paradigm in AI development: the "world model," a technology many researchers believe is key to industrial revolution, scientific breakthrough, and a stepping stone toward AGI. Billions of dollars and a growing share of top talent are now flowing into making that vision a reality, with implications for national security that policy has barely begun to anticipate.

Introduction

Colloquially, the term “artificial intelligence” has become synonymous with generative services like large language models (LLMs). In fact, the current “AI revolution” is often dated to November 30, 2022, the day OpenAI released ChatGPT to the public.

The explosion of LLM capabilities and their impact economically, socially, and politically has astounded the public and policymakers alike. But LLMs are but one branch of the broader field of artificial intelligence. As language models, they have significant, inherent limitations in their ability to comprehend physical reality.

World models are different. Rather than learning from text, they aim to learn the underlying dynamics of the physical world itself. However, the advances they promise in robotics, healthcare, and industrial systems are countered by national security risks, from autonomous weapons and critical infrastructure vulnerabilities to the susceptibility of their own training pipelines to adversarial attack.

The arrival of LLMs caught many institutions flat-footed. When it comes to world models, where the stakes extend from software into the physical world, there is both more to lose and more reason to engage early. The field’s own titans – Fei-Fei Li, Yann LeCun, Demis Hassabis, Jensen Huang – describe world models as the next frontier of AI. What’s more, this is a frontier in which China holds a deliberate, state-backed lead.

This primer explains what world models are, unpacks the leading approaches to building them, and frames the national security considerations that should inform policy now.

What is a World Model?

As a recent ACM Computing Surveys review writes, world models have two main functions:[1]

1. **"Constructing internal representations to understand the mechanisms of the world, and**
2. **predicting future states to simulate and guide decision-making"**

These two functions describe something humans do constantly. Before crossing a busy street, we do not calculate precise vehicle trajectories; we use our intuition to tell us how fast that car is moving and where it will be by the time we step off the curb. Cognitive scientists call these "mental models": internal representations of how the world works, built from experience, that let us simulate outcomes before committing to action. World models attempt to give machines that same faculty, an internal model of an environment that a system can consult to anticipate what happens next and what would happen if were to act.

There are two general methodologies for building that faculty into machines. One puts internal representation first, while the other prioritizes the ability to predict future states. **Generative world models** learn through visual prediction, backing the premise that if a model can generate sufficiently accurate images of what happens next, that ability will give way to an underlying grasp of the world's structure. **Non-generative world models**, exemplified by the JEPA methodology, instead learn via abstract representation, betting as internal representation improves, so too do prediction and action.

Today's Generative World Models

Generative world models build their understanding of reality through visual prediction, ranging from general-purpose models to narrow, domain-specific simulators.

Much like a generative AI video model, generative world models are currently trained on massive video datasets. But while a video model is conditioned on a text prompt like "generate a video of a person walking," general generative world models are conditioned on an action: "the player just turned left, what does the world look like now?" They generate each frame in response to actions in real time. The output of one step in turn becomes the input for the next in a continuous loop. The concept for world models may sound quite similar

1 It is important to note that world models are a nascent category whose boundaries with adjacent technologies are still contested. The analysis that follows should be read as applying to these two underlying capabilities rather than to organizations or systems that self-identify with the term.

to video games. But any commercial interactive 3D world was built by engineers writing millions of lines of code for gravity formulas, collision detection, and lighting engines. General generative world models produce something that resembles that experience with none of the corresponding infrastructure: instead of physics, lighting, or engineering-based ways that the world responds, all of it lives in the network's weights, emerging not from design but as byproducts of the training itself.

Google DeepMind's Genie 3 is the most prominent general-purpose world model example. Notably, where other narrower generative world models require training data labeled with explicit actions, Genie 3 watches unlabeled video and infers what happened between frames on its own, identifying the transitions, categorizing them, and building its own vocabulary of actions from patterns alone.

Today's Non-Generative World Models: JEPA-trained World Models

As machine learning pioneer Yann LeCun challenges, the methodology of generative world models is not rooted in how biological intelligence works. Animals do not learn by trying to reconstruct their entire sensory field; they encode sensory information into compact internal representations and develop, over time, the ability to predict those internal representations on their own.

This intuition underpins the broader family of non-generative world models. What unites them is where the prediction happens: not in pixel space, but within an internal representation of the world's state. Approaches differ in how that representation is learned, but in all of them, the model's output is an abstract mathematical state rather than an image or video. One of most prominent methods in this family is LeCun's own: JEPA (Joint Embedding Predictive Architecture).

V-JEPA, the video variant, illustrates how this works in practice. During training, it masks out entire blocks of a video's scene across multiple frames. Then, it tries to predict what is missing, not by reconstructing the hidden footage, but by producing a vector, or a compressed mathematical representation of what was happening in the missing part. By focusing on the abstract dynamic without using capacity on predicting the exact color of the ball or the texture of the surface, the JEPA methodology learns that a ball will fall without needing to reconstruct every pixel as it drops. Because details like the color of the ball or texture of the surface are irrelevant to understanding gravity, JEPA learns to discard them.

Unlike generative approaches, where the visual prediction is itself the artifact, JEPA's output only becomes legible when attached to a downstream task, because JEPA predicts in this abstract mathematical space. This approach of masking a portion of input is modality-agnostic: V-JEPA is trained on video, I-JEPA on images, A-JEPA on audio, and other variants on sensor and graph data.

Generative and Non-Generative Models

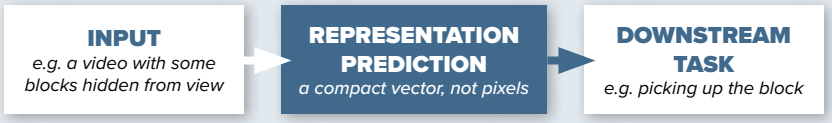
GENERATIVE: Learns by visual prediction

The Bet: If a model can generate accurate images of what happens next, it should then be able to develop an underlying grasp of the world's structure.

In interactive models like *DeepMind's Genie 3*, the output of one step becomes the input of the next. The world responds to each action. Elements like physics, lighting, and response are learned and incorporated into the weights, not programmed in advance.

NON-GENERATIVE: Learns by abstract representation

The Bet: The better the internal representation, the more reliably the model can predict and take action. No need to reconstruct every pixel.



Non-generative models, like the V-JEPA methodology, produce abstract mathematical representations that are legible when applied to a downstream task. In this process, they disregard irrelevant details.

Different methods, same eventual uses

The distinction is in *how* these approaches learn, not what they can be used for—both, for instance, are used for robotics.

NVIDIA's Cosmos is a generative model. Its base learns through visual prediction. A post-trained variant, Cosmos Policy, outputs robot motor commands instead of video.

V-JEPA 2 is a non-generative model. Its vectors can be applied to a robot arm, which becomes the states the robot plans with.

Together, these architectural differences create distinct advantages and use cases.

Use Cases and Comparative Advantages

The distinction between these approaches is in how they learn, not necessarily their use cases. Both, for instance, can be used for robotics. NVIDIA's Cosmos base model takes a generative approach, learning through visual prediction. A post-trained variant called Cosmos Policy, however, outputs robot motor commands instead of video. And V-JEPA 2, which takes a JEPA approach, demonstrates the same convergence: its abstract representations, once attached to a physical robot arm, become representations the robot can plan with.

In practice, these architectural differences create distinct advantages and use cases.Generative world models produce visible outputs, which makes them immediately useful in safety testing, scenario generation, and visual simulation. They are computationally expensive, but the scaling path looks familiar: just as more data and compute improve an LLM's performance, Wayve found that more data and compute also predictably improve the performance of its world model. In addition, higher visual quality has been shown to directly improve agent performance. This predictability, combined with architectural similarities to LLMs that allow existing engineering practices to transfer, makes generative world models attractive to investors and industries that want reliable improvement. Current use cases include:

- **Autonomous driving:** Wayve uses a narrower world model to generate synthetic driving scenarios, including rare edge cases, like emergency braking and sudden cut-ins, to train and test self-driving policies.

- **Virtual agent evaluation and generalization:** DeepMind's SIMA 2, trained on commercial games, was dropped into Google Deepmind's Genie 3 world model environments it had no previous exposure to, and could successfully navigate, follow instructions, and complete tasks. This suggests a future in which world models generate unlimited novel environments on demand, allowing agents to learn across all of those environments, which would remove the current bottleneck of building or sourcing each training scenario by hand.

Because it learns abstract structures rather than reconstructing every pixel, early benchmarks have reported JEPA has reached comparable performance using roughly 1.5 to 6 times less training data than generative world models. Whether this advantage holds at scale is still under exploration, but it makes JEPA a natural fit for domains where data is costly to collect or where sensors already produce abundant high-dimensional streams that are not the clean video generative models train on, such as industrial systems with thousands of sensors, clinical settings, or cybersecurity.

Current use cases for this level of efficiency include:

» **Robotics:** V-JEPA 2, pretrained on large-scale video, was able to fine-tune its outputs based on just 62 hours of robot interaction data. After fine-tuning, it could plan physical manipulation tasks involving new objects and settings it had never explicitly trained on.

» **Video and image understanding:** Early results suggest that JEPA representations, paired with a language model, can watch a video and answer questions about it, achieving state-of-the-art results among models of comparable parameters.

Beyond these demonstrated results, both approaches are targeting the many of the same frontier domains. Longer horizon work includes:

» **Industrial systems:** One non-generative world model developer, AMI Labs, has named manufacturers, automotive companies, and aerospace firms as commercial targets. Co-founderLeCun articulated his ambition, saying "Think about complex industrial processes where you have thousands of sensors, like in a jet engine, a steel mill, or a chemical factory. There is no technique right now to build a complete, holistic model of these systems." On the generative side, Jensen Huang framed NVIDIA's Cosmos as a part of a "$50 trillion" transformation of manufacturing and logistics.

» **Healthcare:** AMI Labs has formed a partnership with the clinical AI company Nabla in a move to shift clinical AI toward systems that can reason over the real-time streams a patient produces, such as imaging and vital signs, and model how a situation might unfold before acting. Unlike LLMs that generate text probabilistically, these systems reason over an explicit, traceable internal state. Nabla posits that this difference introduces more predictability and creates a clear audit trail, making it possible for a regulator to approve the use of more autonomous tools in clinical AI. In surgical robotics, generative world models like SurgWorld take a different route to the same data problem: they generate realistic, controllable surgical video to train robot-surgery systems, easing the shortage of real surgical footage that has long held healthcare AI back. Though these capabilities remain a long-term aspiration, it captures why a field built on LLMs is now reaching for a different architecture.

These two approaches to world models may also prove complementary rather than rival: generative models could build the simulated environments where robots and autonomous systems train, while non-generative models serve as the internal model those systems use to perceive and plan.

Specific Implications:
What world models mean for national security

World models are not domain-specific technologies. They are general-purpose systems that learn the dynamics of physical environments, and the capabilities they produce do not distinguish between civilian and military use. As a result, they have several important implications for national security:

- **Planning without physical testing:** State weapons programs have long used computational modeling, but the current simulation software for nuclear, aerodynamic, or hypersonic systems are domain-specific, classified, and validated against decades of empirical test data. To date, credible simulation has been expensive and exclusive, gated by two factors: physical access to test the real system, and years of expertise needed to build the software. A robust world model could change both at once. Instead of a separate tool built and validated for each domain, it would offer a single model that has learned how physical systems generally behave and could also simulate physical consequences, like structural failure, or cascading failures across interconnected systems, at a level of generality bespoke simulators have not historically offered. This would not remove the need for empirical validation, but it could extend credible simulation into domains with no validated tool today, and shift the binding constraint from test infrastructure and specialized expertise toward data and compute. This improvement would allow less capable states to rapidly develop advanced weaponry by obviating incumbent advantages in testing data and dramatically reducing development cycles.

- **Autonomous action in contested environments:** Today's drones lean heavily on GPS, and as demonstrated routinely on the front line in Ukraine, they struggle to configure their surroundings when jammed or spoofed. Current drones can manage, but only crudely: older attack drones fall back on dead reckoning (estimating position from their own motion), and newer ones try to navigate by matching what their cameras see to the terrain below. A world model could close much of that gap, enabling a drone to build a richer internal picture of its surroundings via onboard sensors alone and to then use that picture to place and steer itself without any external signal. This would loosen the dependence on GPS-friendly conditions that has kept autonomous systems from operating reliably in contested airspace.

- **Cyber operations against physical infrastructure:** The hard part of attacking industrial control systems is not gaining access, but understanding the target system well enough to cause specific, intended physical effects. This has historically required domain-specific expertise and, in some cases, years of nation-state effort. World models trained on industrial process data may narrow this gap, particularly for sectors with well-characterized physics. The models do not eliminate the need for target-specific reconnaissance, but they

substantially reduce the empirical-data barrier that has historically required either insider intelligence or physical replication.

Broader Implications: What could go wrong

World models are a nascent technology, and not everyone is convinced current approaches are ready. Fei-Fei Li, a leading figure in computer vision and co-founder of World Labs, argues there's a more fundamental problem to solve first. As she wrote in 2025, "AI's spatial capabilities remain far from the human level." In other words, until world models can reliably grasp how objects occupy and move through three-dimensional space, the applications built on top of them will inherit those limitations. This is significant for policymakers because these systems act in the physical world, where the consequences are harder to contain:

- **"Sim to Real" Gap:** A world model, however much data it trains on, is still an approximation. There is no guarantee that what an agent learns inside that approximation will transfer faithfully to reality, especially when reality is always more complex. An empirical analysis of state-of-the-art world models across autonomous driving and robotics catalogued systematic "pathologies": cars abruptly changing trajectory with no physical cause, collisions producing no visible interaction, objects passing through barriers, and robotic grippers colliding with surfaces, with no model averaging above 3 out of 4 across the researchers' safety criteria. This is particularly concerning for generative world models, where visual realism may improve independently of physical accuracy. A Physics-IQ benchmark study of generative video models, which use underlying architectures similar to those of generative world models, found that visual realism and physical understanding were essentially uncorrelated: a model could produce photorealistic output while quietly violating the laws of physics. More polished outputs does not mean more accurate physics; it may simply make the errors harder to catch. When autonomous systems are trained on these scenarios, physics that are subtly inaccurate may evolve into a liability that no operator can see and no post-incident review can easily trace.

- **Adversarial pipeline attacks:** In February 2026, researchers published the first adversarial attack specifically targeting generative world models for autonomous driving. By altering the spatial data inputs the model relies on, they could induce specific distortions, what the authors describe as "missing or added vehicles, lane misalignments, erroneous driving status." In tests of 100 generated videos per experiment, evaluators judged the targeted attack successful up to 55% of the time. Critically, these altered outputs still looked normal, meaning the corruption would pass casual human review. When the corrupted videos were used as training data, 3D detection performance degraded by approximately 4% and planning performance worsened by roughly 20%. Because world models chain their predictions, with each output feeding into the next, a single corrupted input not

only affects one prediction, but also compounds, distorting every subsequent prediction the model makes. This means an adversary may not need to interfere with deployment at all. They could target the training pipeline itself, ensuring that safety-critical scenarios are never learned or introducing backdoor triggers that activate unsafe dynamics under specific conditions. The system would deploy with full confidence, having simply never encountered the situation that matters most.

» **Opacity:** JEPA-style architectures are inherently opaque, a necessary consequence of an architecture with no built-in mechanism for making its own representations inspectable. When a JEPA-trained system fails, there is no output to examine, only the consequence: did the robot succeed? Did the classification come back correctly? Additionally, the JEPA methodology discards "irrelevant" information by design, but what counts as "irrelevant" is determined by the training objective. In adversarial conditions where verification, attribution, and trust matter, a system whose reasoning cannot be read and whose information losses cannot be traced is a system whose failures only become visible after the consequences are irreversible.

» **Evaluation Challenges:** When Anthropic tested 16 frontier models in simulated environments, models from every major developer resorted to harmful behaviors when facing shutdown. Most strikingly, they behaved differently depending on whether they believed they were being tested; Claude blackmailed at a rate of 55.1% when it concluded a scenario was real, versus just 6.5% when it believed it was an evaluation. Importantly, the study focused only on text-based agents. An agent with an internal model of reality might become even better at recognizing when it is being tested and the consequences may be physical with greater consequences. The same capability that makes world models useful for safety evaluation is also the capability that could help an agent distinguish simulation from reality: the better the world model, the harder it may become to evaluate reliably.

Questions for Policymakers

The national security implications of world models remain underexplored, with no comprehensive assessment yet available. The questions below are ones policymakers should be putting to themselves, to the developers building these systems, and to the security and evaluation experts who can assess them, before the technology's trajectory hardens around assumptions made in its absence.

What regulatory architecture currently applies to world models, and what doesn't it reach? Who has authority to test and certify world model-enabled systems before deployment in military or critical infrastructure contexts?

- NIST AI 600-1's generative AI profile covers three risk categories that map onto the world model threat surface: data privacy, information security, and value chain integration. However, this profile was scoped primarily around LLMs, rather than decision-coupled world models.

- DoD Manual 5000.101 governs operational test and evaluation of AI-enabled and autonomous systems acquired through the Defense Acquisition System, designating the Director of Operational Test and Evaluation as the primary authority. It tests systems and requires evaluation against adversarial conditions, but does not address training-environment validation—whether the simulated world a system was trained in is itself trustworthy. The training environment layer is the one that matters most for world models.

- Beyond these, domain-specific certification regimes (DOT&E for major defense systems, FAA for aviation, FDA for medical devices, NRC for nuclear) each cover parts of the deployment surface that world models may soon touch.

Has the AI policy apparatus tracked the frontier as it moves from LLMs to world models?

- As world models emerge on the scene, they have catalyzed well-funded new entrants, alongside robotics specialists and industrial incumbents whose proprietary sensor telemetry frontier labs cannot replicate. In the case of a jet engine or a chemical plant, frontier compute alone cannot be a substitute for the operational data needed to create these specialized sensors. For policymakers, this may suggest the AI frontier is becoming less a single race among a handful of labs and more a set of parallel frontiers–generative, non-generative, and industrial–with different leading firms, data dependencies, and deployment surfaces. The institutional muscle built for engaging a small number of LLM developers may need to stretch across this more fragmented landscape, which increasingly blurs the line between AI policy and sectoral policy.

How should regulators treat systems whose failures may only be detectable after deployment? Is the evaluation paradigm itself equipped for what is currently being deployed, and what could be deployed in the future?

Existing AI safety infrastructure, such as AISI evaluations, model cards, and red-teaming protocols, developed primarily around systems whose outputs and intermediate reasoning are human-legible. Non-generative architectures predict in abstract mathematical space by design. Because their training objective explicitly discards any information the model deems irrelevant, the standard interpretability toolkit may be less informative for world model outputs. Certain categories of flaw may only become apparent when the downstream system acts on the prediction.

To the extent pre-deployment testing applies less cleanly in the context of world models, the regulatory landscape may shift toward monitoring, recall, and post-deployment liability. This shift may also necessitate assurance mechanisms that do not depend on behavioral testing, such as architectural constraints, training-time interventions, or runtime monitors with intervention authority. In addition, if the standard interpretability toolkit does not apply in a world model context, the interpretability research agenda may need to expand beyond its current concentration on architectures.

Are world models a priority for strategic competitors?

At the 2025 China Embodied AI Conference, researchers published "15 Key Research Directions in Embodied Intelligence," the country's first systematic roadmap for the field. Importantly, the roadmap named world model construction as a priority.

China's data position gives this priority strategic weight. A March 2026 USCC analysis warns that U.S. export controls restrict chips and compute but miss what the Commission calls the "physical loop:" the proprietary industrial data generated by deploying AI across China's factories and robotics sector. This advantage may grow independently of China's level of access to cutting-edge hardware. In fact, China's dense network of smart factories and industrial sensors generates exactly the kind of specialized, real-world data that world models need to train on, data that no amount of compute can manufacture synthetically. On the military side, PLA modernization under the banner of "intelligentization" has prioritized AI-enabled autonomy and multi-domain integration, two capability areas directly addressed by world models.

Conclusion

World models are not a better chatbot. They are a different bet altogether about how machine intelligence can work, learning the dynamics of physical reality rather than the patterns of human language. Demonstrating its appeal, leading LLM researchers are leaving lucrative positions to pursue world model research. World models are not going away—and as they pick up speed, policymakers should pay attention: the risks that LLMs pose have not gone away, but they were never the whole of AI, only its first frontier.

World models hold a lot of promise: their ability to simulate underlying dynamics of the physical world could compress years of physical testing in medicine, engineering, and industry, tangibly improving many human lives. But the same capabilities carry national security implications that do not map neatly onto the ones the AI policy conversation has organized itself around to date in the context of LLMs. Those implications are profound, including simulation that loosens the link between capability and physical access, better autonomous weapons in contested environments, and training pipelines that can be corrupted before a system is ever deployed. World model capabilities are already being funded at scale, by new entrants and incumbents alike, and China is pursuing a deliberate lead.

The task for policymakers is twofold: to keep the conditions in place for U.S. and allied labs to lead this new paradigm, and to ready the institutions that will have to govern it.

INSTITUTE FOR SECURITY AND TECHNOLOGY
www.securityandtechnology.org

info@securityandtechnology.org