Stanford University
Human-Centered
Artificial Intelligence

Issue Brief
HAI Policy & Society
July 2026

The World Model and Spatial Intelligence Era: Governing AI Beyond Language

Daniel Zhang,* Russell Wald,* Ehsan Adeli, Elena Cryst, Daniel E. Ho, Caroline Meinhardt, Jiajun Wu, Amy Zegart, Li Fei-Fei

Introduction

It is 2 a.m. and a wildfire is rapidly moving toward a densely populated suburb. An incident commander is trying to answer several questions at once: which evacuation routes are still passable, which hospitals are reachable, and where the power grid is likely to fail next. Current incident support systems, including AI-enabled modeling and physics-based simulations, provide real-time forecasts. But these projections remain siloed, and conditions change faster than analysts can track.

A world model — an AI system that builds and maintains a working representation of an environment to predict how conditions may change — addresses both gaps. In this case, it could bring forecasts about roads, the power grid, and hospital capacity into one continuously updated picture. The commander could use that picture to reroute crews, redirect resources, and test evacuation plans before issuing an order.

Key Takeaways

World models are AI systems that build a working representation of an environment to predict how it changes in response to action. They could lower the cost of high-quality simulation, benefiting infrastructure planning, crisis response, experimentation, and embodied AI training.

No existing benchmark gives policymakers an adequate basis to evaluate a world model for safety-critical deployment. Closing that gap requires public investment in measurement science.

Policy built for existing AI-generated content and autonomous decision-making does not fully address the risk profile of world models. The distinctive question is whether a simulated environment matches physical reality closely enough to train or test another system or guide a real-world decision.

The scarcest input is action-labeled interaction data — robot trajectories and fleet logs that cannot be scraped from the web, which risks concentrated control. Public datasets should be an explicit target of federally funded research.

World models are dual use, with national security implications. By lowering the cost of capable autonomous systems, they could open military advantage to less-resourced entrants, making early leadership in world-model research, development, and governance an urgent national security priority.

* Equal Contribution

The emergence of world models points to a broader shift in AI. The systems that define the current era operate primarily through language. Multimodal systems expand that capability to images, audio, and video, but most still cannot track a coherent environment over time. World models, by contrast, aim to maintain a coherent representation of an environment over time and predict how it may change in response to action. That capacity — to understand a physical environment and use that information to guide action — is spatial intelligence. World models are one technical pathway toward achieving that capability.

World models raise a distinct governance challenge: They can serve as stand-ins for the physical world.

Beyond crisis response, world models can help manufacturers test designs before building them, engineers assess infrastructure under stress, and robots work in settings that are too dangerous for people. But moving AI closer to the physical world raises new policy concerns, such as flawed simulation, pervasive sensing and privacy erosion, questions of liability, concentrated control of spatial data and infrastructure, and dual-use risks in national security.

World models demand a broader policy agenda than the current wave of governance discussions around language-centered AI. While many applications will intersect with existing governance regimes — such as those for autonomous vehicles, medical devices, industrial robotics, infrastructure planning, and defense systems — world models add a capability layer that cuts across these sectors and demands attention across multiple domains. World models also raise a distinct governance challenge: They can serve as stand-ins for the physical world. Policymakers must therefore ask whether a learned environment matches reality closely enough to train a system on, test a system against, or guide a real-world decision.

For policymakers, the goal is twofold: to enable the socially valuable uses of world models while guarding against flawed simulation, concentrated control, and unsafe deployment. This brief sets out three governance priorities: broad access to the technology, safeguards matched to how and where a system is used, and the public capacity to evaluate these systems independently.

World Models: An Introduction

The idea of a world model predates the term "artificial intelligence." In 1943, psychologist Kenneth Craik proposed that human beings carry a "small-scale model" of reality to test actions internally before acting in the real world — for example, reasoning through whether a bridge design would work in our minds before building it. Early AI researchers took up this concept in the 1960s and experimented with giving machines structured, internal models of how the world works, though for decades, such systems remained restricted to narrow environments governed by explicit rules.

Several recent developments changed that. First, advances in reinforcement learning showed that agents could learn to act through internal simulation, reducing

their reliance on costly real-world trial and error. Second, the rise of multimodal AI made it possible to build AI systems grounded in how the physical world actually looks, moves, and sounds. Together, these advances paved the way for modern world models.

A world model learns how an environment behaves from large amounts of perceptual and interaction data — images, video, sound, sensor recordings, and records of actions and their outcomes — just as a person learns how the world works by observing it closely. A central goal of world models is counterfactual reasoning: predicting the consequences of an action, including actions the model was never explicitly trained on. This is a key distinction between a world model and a language model. A language model predicts what comes next in text; a world model predicts how an environment may change in response to an action, simulating what happens if an object is moved or a door opened.

In practice, the ability to generate coherent environments that respond to action remains at an early stage, but it is advancing quickly. Early models could generate short, realistic clips but soon lost track of the scene, as objects flickered, drifted, or disappeared. Some newer models maintain consistency for longer and allow users to act within the generated environment. One test of further progress is whether an agent can move an object, leave the scene, and return to find it where it was left. This consistency, when it holds, is learned from observation rather than built in by hand, rule by rule, as in conventional simulation and video games, where every object is placed in advance.

The field is moving toward models that generate more interactive worlds, maintain coherence over longer stretches, and generalize to environments not seen during training. Progress depends on both computing power and data that captures how the physical world responds to action, which is far scarcer than the text and images that trained earlier systems.

With growing funding in this space, tech incumbents on both sides of the Pacific, including Google DeepMind, Nvidia, Alibaba, and Tencent, as well as newer startups such as World Labs,[1] AMI Labs, and Odyssey, are leading this world-model push. Recent commercial systems can now produce narrow, interactive 3D environments from text or images. However, a general-purpose world model capable of predicting physical consequences and supporting reliable planning and action across diverse real-world environments remains an active research frontier.

A Functional Taxonomy of World Models

Diverse approaches are important for advancing the field. Researchers are building world models using a range of technical architectures (see examples of labs exploring different approaches), but for the purposes of this brief, policymakers can better understand them by the role they play in the broader context of perceiving, understanding, and acting in the physical world. Brief co-author and Stanford Professor Fei-Fei Li proposed a functional taxonomy of three categories: renderers to help us see, simulators to help us understand, and planners to help machines act. This approach views a world model as a stack of capabilities — rendering what a world looks like, simulating how it behaves, and planning how to act within it.

1 Fei-Fei Li is currently on partial leave from Stanford University and serves as the co-founder and chief executive officer of World Labs

Table 1. Core functional categories of world models

CATEGORY	CORE TASK	PRIMARY OUTPUTS	EXAMPLE APPLICATION AREAS
Renderer	Generate observations of a world from a viewpoint, prompt, or scene descriptions.	Images, video, visual sequences, interactive views.	Architecture and design visualization; animated film production; digital content production.
Simulator	Model the underlying state, geometry, and dynamics of a world; show the effect of changes in a given environment.	Digital environments that behave according to physics and biological properties.	Crash test simulator; digital twins; practice environment for medical surgery; military training in air, land, sea, and space domains.
Planner	Determine what action an agent should take given a state or goal.	Actions, trajectories, policies, intervention choices.	Robotics, autonomous vehicles, warehouse systems, healthcare logistics, disaster response, defense and security applications.

Note: This framework is borrowed from a blog post titled "A Functional Taxonomy of World Models," written by brief coauthor Fei-Fei Li and published on the World Labs Substack on June 3, 2026.

Renderers generate observations, typically images or video, from a particular viewpoint or prompt. This makes renderers valuable for design, architecture, immersive media, and visualization. A renderer can generate a photorealistic image of a new hospital wing to help architects visualize the space before construction begins. However, renderers are optimized for plausibility rather than underlying truth. Their output may look convincing without preserving stable geometry or physical consistency. In the hospital example, the model may not capture how people, materials, or infrastructure would behave within that environment.

Of the three categories, renderers are the most commercially mature: Interactive generators like World Labs' Marble and Tencent's HY-World 2.0 already produce explorable scenes from text or image prompts.

Simulators model the underlying state of the world — its geometry, physics, material properties, and dynamics. Beyond making a world appear realistic, a simulator makes it behave in ways that are structurally consistent and computationally useful — qualities that make simulation valuable for engineering, robotics, digital twins, scientific research, and industrial design. A simulator can test how a robot behaves in a warehouse, how a molecular system evolves, or whether a bridge could bear the weight of a military convoy. Explicit simulation may be an important path to physical validity, but it is not the only one: Recent unified world-action models learn rendering, dynamics, and control jointly, with planning capability emerging directly from learned visual dynamics rather than from an intermediate simulator.

Conventional simulation is already mature: Factory and city models run on platforms like Nvidia's Omniverse, but building each environment is slow, specialist work. Because a world model learns its simulation from data, new environments can be created more quickly and updated as conditions change. This approach is still early, limited by scarce physical data.

Planners generate actions. Given an observation and a goal, they determine what an agent should do next and what the effect of that action will be — making planners the bridge from world understanding to world intervention. They are essential for robotics,

autonomous vehicles, and other systems that must operate in unstructured environments. In a disaster zone, a planner can guide a robot through a damaged building as debris and people shift around it, turning what it sees into real-time actions.

Planners are the least mature of the three categories. Some autonomous vehicles and warehouse robots can operate reliably within bounded domains, but most contemporary robot demonstrations are confined to short, narrow tasks in controlled labs. Generalizing these capabilities to unstructured, dynamic environments remains an unsolved challenge.

Capability thresholds defined per category are easily gamed or outgrown; consequently, safeguards must attach to the deployment context rather than to the model class.

When these functions operate together in a real-time, action-conditioned loop, a further capability emerges: interactivity. People and machines can act within a modeled world and use the resulting feedback to guide action in the physical one. A surgeon in training could rehearse procedures inside a simulated anatomy that responds to each movement; a warehouse robot could practice handling unfamiliar objects in a digital twin before working in the facility itself; and military units could rehearse remotely with allies and partners, lowering costs while limiting danger to personnel and exposure to adversary surveillance. Interactivity is thus a natural consequence of a capable world model and where its most valuable uses lie.

The boundaries between these categories are already blurring as renderers become interactive, simulators more generative, and planners more capable of reasoning. At the research frontier, unified models increasingly combine rendering, simulation, and control within a single network. For policy, this matters: Capability thresholds defined per category are easily gamed or outgrown; consequently, safeguards must attach to the deployment context rather than to the model class.

Why World Models Matter

Widely available world models could have tremendous impact on efficiency and innovation. With the potential to reduce the cost of and increase access to high-quality simulations, thereby enabling novel experimentation by small and large firms, world models could help improve infrastructure planning, crisis response, and scientific discovery.

Better Simulation and Public Planning

The most immediate public benefit is better simulation. High-quality simulation is already essential in engineering, logistics, and infrastructure planning, but accurate models remain expensive to build and update. Modeling a factory, hospital, or city district can require substantial specialist work to encode its geometry, laws of physics, and changing conditions. World models could lower that barrier, making simulations easier to create and adapt.

For the public sector, world models offer a digital twin — a working approximation of a city or system that can be used for infrastructure and emergency

planning. A city could use such a model to examine how infrastructure disruptions might cascade across systems and compare response options before a crisis. Sector-specific crisis-modeling tools already simulate many of these risks but are difficult to combine into a coherent picture of a fast-moving event. World models could integrate these risks into cross-domain simulations, providing sharper foresight and better targeting of limited public resources.

Lowering the Cost of Physical Experimentation

Because world models could lower the cost of creating high-fidelity simulation environments, they could also lower the cost of experimentation in physical industries. An aerospace firm could test how a new wing performs under various conditions before committing to a physical prototype; a hospital could model patient and staff flow through a redesigned emergency department during periods of low and peak demand. As the technology matures, the same approach may open new avenues for experimentation in less well-known domains, including maritime and space operations. The result is fewer costly real-world mistakes along the path to a workable design.

Training Grounds for Embodied AI

Another benefit is in embodied AI — systems that sense and act in the physical world, such as robots, drones, and self-driving vehicles. Because real-world testing is costly, dangerous, or too rare to capture in data, training robots in simulation before real deployment is already standard practice. The open challenge is transferring that training to messy and dynamic real-world environments.

For example, a robot in a chemical plant may need to cross a leak zone where visibility is poor and turning the wrong valve would result in serious consequences. Or a maritime drone may need to navigate contested waters, avoiding vessels and underwater hazards to rescue pilots after an aircraft goes down. In each case, a machine that can model a changing environment and predict the effects of its actions accomplishes far more than one that follows a script. World models' long-term value lies in helping machines operate safely.

A Broader Innovation Layer

World models could become shared infrastructure that smaller players build on. Currently, the cost of high-quality, domain-specific simulators can price out startups and public sector organizations. A common world-model layer could give those organizations access to advanced simulation without requiring each one to build the underlying models and infrastructure, much as cloud services provide computing capacity without requiring users to own data centers. This could put advanced simulation within reach of a university lab or a city agency. Broader access allows more entrepreneurs, researchers, and public institutions to develop competing applications — a potential advantage for the United States rooted in its strengths in entrepreneurship and fundamental research.

Risks and Policy Challenges on the Horizon

World models enter a policy landscape that earlier AI technologies have already shaped. Computer vision forced a reckoning over surveillance and facial recognition; robotics and autonomous vehicles prompted regimes for physical safety, certification, and liability; foundation models surfaced risks from discriminatory outputs, deception, unauthorized use of copyrighted material, and malicious applications of dual-use capabilities.

Much of what world models raise is already familiar: Multimodal models cut across sectors, and agentic systems act without a person deciding each step.

Rules for outputs and permissions for action are therefore insufficient on their own; governance must also establish whether the learned environment is valid for its intended use.

The more distinctive issue is what else must be governed. Foundation-model governance has often centered on content — what a system generates. Agentic AI governance focuses on the authority to act — which actions a system may take and under whose oversight. World models bring a third object into focus: the validity of the learned environment itself. A world model's error is a counterfeit of physical reality that can look flawless while being wrong — and every system trained inside it, every certification that relies on it, and every decision informed by it inherits the flaw silently, at scale. Rules for outputs and permissions for action are therefore insufficient on their own; governance must also establish whether the learned environment is valid for its intended use.

The Visual Plausibility Trap

A renderer can produce a persuasive depiction of a world without modeling the geometry, dynamics, or physical constraints needed for safety and reliability. An AI-generated video may depict fire or flowing water without obeying the laws of physics. A generated building can look sound without a stable underlying structure.

The trap varies across model designs. Video generators without a persistent scene representation may lose consistency over time. 3D-native systems use explicit geometry to improve spatial consistency but still fail to capture how a world changes. Models that use a compressed internal representation of the environment, known as latent state-space models, prioritize predicting change over visual detail. Each approach fails differently and therefore demands a different evaluation. Even at the frontier, the limits are concrete. Genie 3, for example, can generate an explorable scene in real time, but at its 2025 release, the world stayed coherent for only a few minutes before objects began to shift or vanish.

The lesson for policymakers is not to confuse visual realism with functional reliability. Certification and procurement for high-stakes uses must therefore rely on testing that shows whether a system's physics hold up and whether it fails safely at the edges, rather than on convincing demonstrations.

The Simulation-to-Reality Gap

The gap between simulation and reality grows critical as systems tested in simulation move toward real-world deployment. Unlike the plausibility trap, where a system looks competent to a human observer, a system may perform well inside a generated or learned environment but fail in the real world because of sensor noise, lighting, weather, wear, unexpected human behavior, or edge-case physics. For example, autonomous vehicles may perform well in a controlled environment, yet struggle with foul weather, glare, or unusual road events their training did not cover adequately.

World models add a failure mode of their own when a learned model is used to train a system and to judge it — the equivalent of teaching to a flawed test. If the model understates the risk of skidding in rain, a vehicle

trained in that model may learn to drive too fast and still score well when the same flawed model is used to test it. The score would reflect an error in the model, not readiness for a real road.

Policy must define how much real-world validation a system requires before deployment, regardless of its performance in simulation. Because even strong tests will inevitably miss certain rare conditions, oversight must continue after deployment, through monitoring and incident reporting that feed failures back into the standard.

Evaluation and Standards

World models require different evaluation questions, yet there are no settled standards for answering them. For language models, evaluators ask whether an answer is accurate, unbiased, safe, or useful. The same questions still apply, alongside others: whether objects remain stable as the viewpoint moves, whether scenes change in physically plausible ways, and whether systems fail safely when they encounter a rare event.

Evaluation should match the system's function. A renderer used for concept art can be judged by how convincing it looks. A simulator used for infrastructure planning must meet a higher bar — whether its geometry and physics are actually correct. A planner embedded in a robot must be tested repeatedly across varied, real-world conditions. The closer a system comes to real-world actions, the more its evaluation should weigh physical validity, robustness, and transfer beyond the test setting. Interactive systems add another question: Do skills practiced in simulation, by a person or a robot, transfer to the real task?

Current benchmarks reflect this gap. Earlier video benchmarks like VBench measure visual quality, prompt alignment, and temporal smoothness, but not whether generated scenes obey physical laws. Newer benchmarks probe that physical understanding: VideoPhy and PhyGenBench test physical commonsense in generated videos; WorldScore assesses controllability, quality, and dynamics in world generation; WorldModelBench judges video models as world models; WorldArena evaluates perceptual quality and usefulness in training, testing, and planning robot behavior; and the related robotics benchmark LIBERO tests how well robots complete simulated manipulation tasks.

Even so, leading models still fail to maintain basic physical consistency, and high benchmark scores can conceal weaknesses in physical reasoning or robustness to changing conditions. Evaluation of world models remains a research patchwork rather than a settled standard — none gives policymakers an adequate basis to assess a world model for safety-critical deployment.

Physical Liability

World models complicate liability because they sit between perception and action. A language model produces words that a person decides how to use, whereas a world model may guide action based on its reading of the environment, often with no human intervention. The problem is not unique: Agentic systems raise the same question whenever AI can act without a human decision-maker. World models bring that question into physical settings, where an agent may act on a learned model's mistaken interpretation or prediction of its environment.

If an autonomous vehicle misreads a construction zone, or a hospital logistics system blocks access to critical equipment, the failure could come from the same black-box causes as a language model — training data, model design, and deployment choices. What world models add is a learned representation of the physical

environment on which an autonomous system may act. That does not eliminate responsibility. It changes what responsibility depends on. Did the deployer use it outside the setting where it was tested? Did the operator have enough information to override it? Did the hardware provider build a sensor configuration that the model could not handle? Existing frameworks have been designed for complex products with several contributors, but world models introduce a learned layer of environmental judgment that makes it more difficult to attribute harm to a particular action or cause.

Spatial Privacy and Pervasive Sensing

World models push AI toward continuous observation of physical environments: more cameras, microphones, mapping, and opportunities to collect data about where people are, how they move, and with whom they interact. Building useful models of homes, workplaces, roads, hospitals, schools, and public spaces often requires sensor-rich data about how those environments are arranged and how people move through them.

The harder problem is inference. A system trained on multimodal spatial data may infer home routines, workplace patterns, health-related behavior, social relationships, or sensitive locations that were never explicitly labeled. Fed by a building's or a city's sensors, a world model can hold a continuously updated picture of who is where and how a space is used over time. As a result, privacy law and procurement policy must address not only raw sensor feeds, but also the downstream creation of persistent spatial profiles and digital representations of people, homes, workplaces, and public spaces. That concern is especially acute in authoritarian regimes, where governments face fewer constraints on surveillance and the uses of information gleaned from it.

Evaluation of world models remains a research patchwork rather than a settled standard — none gives policymakers an adequate basis to assess a world model for safety-critical deployment.

At the same time, world models could bolster privacy where simulation can substitute for real-world monitoring. Autonomous vehicle developers, for example, could rely more on simulated road environments for training and testing, reducing the need to collect comparable footage from public roads. Simulators would still require real-world data to build and validate, but could reduce repeated collection as testing scales.

Concentration and Dependence

World models may increase market dependence on a small number of firms that control the best multimodal spatial data, the largest compute budgets, and the most advanced simulation infrastructure. Their advantage may be durable: Passive visual data is abundant, but action-labeled interaction data — including robot trajectories, teleoperation logs, and fleet sensor streams paired with control signals — is far more difficult to obtain at scale. Tech incumbents that already deploy widely accumulate inputs unavailable to competitors and public institutions, with each deployment compounding their advantage.

For governments, this is not only a matter of market competition but also of strategic capacity — who owns, governs, and can audit the simulation infrastructure on which future autonomous systems may depend. An agency buying a self-driving shuttle or an inspection drone must verify its safety by testing rare and dangerous situations that cannot easily be repeated in reality. Simulation is often the only practical way to do that at scale. But the most useful system-specific simulator is often the vendor's own because it depends on proprietary details about the system's sensors, architecture, operating data, and known failure modes. The agency may be unable to distinguish a robust system from one that performs well only under vendor-selected conditions. Independent evaluation is therefore central to procurement. Without access to the simulator or an independently validated alternative, the agency then depends on the vendor both to build and validate the system.

Countries that control leading world-model infrastructure and data will shape how others access and use these systems. Governments without domestic alternatives may become reliant on foreign providers, limiting their ability to inspect, adapt, or replace systems used in critical functions.

Labor and Physical Work

Earlier waves of AI primarily focused on knowledge work such as writing, analysis, and software development. World models extend automation to physical and operational work. Conventional robots already handle repetitive tasks in controlled settings like assembly lines. World models reach further, into jobs that require judgment in unstructured, changing environments that have resisted automation, such as driving, warehousing, and construction. As world models mature, the impact on the workforce broadens beyond the office and the factory floor.

That shift will also reshape the nature of expertise. As physical operations are increasingly captured in simulation — how a plant runs or how a crane operator executes a lift — that knowledge moves from the workers who hold it toward the firms that build the models. An operator or an agency can gradually lose the ability to perform the work without the system. The policy question then becomes whether workers and public institutions have a role in how these systems enter real workflows, rather than simply receiving them as finished tools.

National Security Implications

World models carry major national security implications. Like other general-purpose technologies, they are dual use: A model that navigates aid through a disaster zone can guide a weapon through the same environment. Modern security competition increasingly depends on the ability to model terrain, rehearse contingencies, and operate with limited communications or human control. The consequences of world models also become more acute in security

The consequences of world models also become more acute in security settings because adversaries could deceive, disrupt, or defeat these systems at machine speed, often under conditions where human oversight is either limited or too slow to matter.

settings because adversaries could deceive, disrupt, or defeat these systems at machine speed, often under conditions where human oversight is either limited or too slow to matter.

Because of their commercial and military potential, world models are likely to become an important arena of future geopolitical competition, particularly between the United States and China. Countries around the world are already aggressively pursuing embodied intelligence and robotics. China, for example, has made them a national priority, identifying embodied intelligence as a strategic industry in its latest Five-Year Plan and directing state funds and shared data infrastructure toward physical AI. China's approach also emphasizes physical-world data and deployment. Whether that can reduce reliance on frontier compute remains uncertain. The United States has led largely through private labs and research on physical AI rather than federal industrial policy; however, in late 2025, the White House was reportedly considering an executive order on robotics, and in 2026, Congress introduced a bill to establish a National Commission on Robotics.

Who moves first in world-model breakthroughs matters. Early deployment can generate proprietary operational data and help set standards, making systems harder to displace across commercial and defense markets. But speed without reliable and auditable systems can create vulnerabilities of its own.

Strategic Advantages

- **Readiness and mission rehearsal.** World models could make synthetic training environments faster to build and adapt, allowing planners to test complex, high-stakes, rare, or dangerous scenarios before they occur. This could shift geopolitical power by enabling countries with limited operational experience to effectively test military plans under unfamiliar combat conditions.

- **Contested operations and logistics.** World models could help planners adapt supply routes and depot operations under stress, while enabling aerial, maritime, and ground robots to carry out logistics missions with degraded communications and intermittent human supervision. Connecting planning with execution could make military sustainment more resilient in contested settings.

- **Intelligence and decision support.** Simulation environments could help intelligence analysts assess adversary actions, trace cascading effects, and better analyze "what if" scenarios. They could also allow leaders to test assumptions before committing resources. At the same time, however, they may make uncertain forecasts appear more precise than they are.

- **Cost and scale.** World models could make drones and robots cheaper and more adaptive across national security missions, including intelligence collection, deterrence, defense, and warfighting. For the United States and its partners and allies, world models could enable large quantities of capable, low-cost systems that could improve defense effectiveness at scale.

Security-Critical Risks

- **New cyber-physical vulnerabilities.** World models inside an operational loop give potential attackers a new target: the system's picture of its surroundings rather than the data it holds. Corrupting that picture can turn a breach into a misguided maneuver or strike. As these systems become more operationally valuable, securing their training data, updates, sensors, and simulation environments will need to advance alongside their development.

- **False readiness.** A synthetic environment may make a platform or plan look more prepared than it is, especially when deception or sensor interference widens the gap between simulation and reality.

- **Increasing risks of miscalculation and escalation.** Better world models may let uncrewed systems navigate and coordinate with little human supervision. But not all countries will draw the same line between human and machine control. Those differences mean that two sides in a conflict could operate at very different speeds and use very different assumptions about what signals mean and what one move or counter-move demands. Because human-autonomy handoffs are hidden for security, neither side may know how the other system operates. These differences could raise the risk of miscalculation and escalation in a crisis.

- **Strategic dependence.** Governments may rely on proprietary simulations for training and operational planning. If those systems cannot be inspected or replaced, critical national capabilities may depend on private infrastructure controlled by a small number of firms, including foreign providers. This creates both sectoral dependence on particular vendors and national dependence on foreign-owned infrastructure.

These advantages and risks expose the limits of current security policy. Export controls have focused on compute and, briefly, model weights on the assumption that these inputs are decisive chokepoints. World-model advantage may also depend on physical-world data and the ability to deploy systems at scale, putting key inputs outside that frame. Interoperability presents a separate challenge. If coalition operations depend on shared simulations, allies trained on incompatible proprietary systems may struggle to operate together. The standards that govern those environments will shape which systems can be combined.

A Governance Framework for the World Model Era

Because world models are an early and fast-moving frontier, responding with a rigid regulatory playbook would be premature and ill-conceived. Instead, the following framework offers strategic guidance designed to anchor policy conversations as the technology matures.

This approach builds on existing governance efforts for language models and frontier AI. Many priorities from the current wave of AI policy apply with equal force to world models. But world models add a capability layer that demands a tailored response. World models' capacity to turn digital representation into physical intervention creates risks that model-level or language-focused governance cannot mitigate. A durable agenda therefore pairs targeted safeguards with active public support, steering world-model development toward public value across its life cycle. That effort consists of three complementary pillars: infrastructure and incentives, proportional safeguards, and public sector capacity.

Infrastructure and Incentives

Early policy should protect the downstream benefits of world models from exclusive capture by a few well-capitalized firms. Since key inputs of such models, including the scarce action-labeled interaction data, cannot simply be scraped from the internet, developers must gather them by operating physical machines in real-world environments. This creates a steep barrier to entry that risks concentrated control over core

A durable agenda therefore pairs targeted safeguards with active public support, steering world-model development toward public value across its life cycle.

simulation infrastructure. While some firms already release datasets and simulation tools, the market may still underprovide shared action data and simulation environments for startups, universities, and public agencies, especially for public-interest applications with diffuse or uncertain commercial returns.

To distribute this capability, governance should broaden access to the underlying computing power and support the development of inputs specific to world-model applications. Shared pools of action-labeled data, such as robot trajectories and teleoperation logs, along with public-interest simulation environments, should be an explicit target of initiatives such as the National AI Research Resource (NAIRR) led by the National Science Foundation (NSF). NSF should coordinate that shared infrastructure, while sector-focused agencies contribute domain-specific testbeds and states support real-world testing through procurement.

These efforts can build on the open ecosystems already emerging from industry and academia, including open world-model weights, training recipes, and evaluation tools. Public funding and procurement can sustain these resources as common reference points and support applications with broad public value but fragmented demand, such as disaster response and infrastructure resilience.

Proportional Safeguards

Governance should track operational function and real-world consequences rather than apply a uniform, blanket regime. The core principle is that the closer a system comes to safety-critical simulation or real-world action, the more stringent the governance should become.

Four elements should anchor early safeguards:

- **Build measurement science:** Existing methods cannot independently confirm that a learned world model performs safely during high-stakes physical deployment. Near-term policy should fund the benchmarks and incident-reporting frameworks that future oversight will need. In the United States, the National Institute of Standards and Technology (NIST) should develop shared evaluation methods, while sector agencies should define relevant operating conditions and reporting requirements. Until those methods mature, existing physical-safety regimes should continue to require rigorous field testing.

- **Ensure independent evaluation:** Ensuring that developers are not the sole judges of their own systems is already a key focus of AI governance. System-level evaluation is especially important for world models because their performance depends on continued interaction with the deployment environment, not only on outputs measured at a single point in time. Independence should extend both to who conducts the evaluation and who defines the test conditions.

- **Protect spatial privacy:** Privacy safeguards must adapt to world models. Data minimization — the principle that only the minimum data necessary to achieve an objective is collected and retained — is difficult for world models. Developing a

simulator is an exploratory process where the relevant data elements may not be known until the model takes shape. This may require what is recognized elsewhere as staged application of data minimization. Once a simulator has been validated, developers should retain raw sensor data only for validation and delete unnecessary elements. Policy should also encourage simulation where it can reduce repeated real-world monitoring, while guarding against its use for persistent physical-world tracking.

- **Document perception and action:** Transparency rules for foundation models already call for documentation and logging; embodied systems need a physical-world version of the same. The requirement is a time-stamped record of what the system perceived, the state it inferred, and the action it took, so a physical incident can be reconstructed and duty of care assigned across developer, deployer, operator, and integrator. Because the internal reasoning of these models remains only partly interpretable, human oversight is imperative where the stakes are highest.

Public Sector Capacity

The tools to evaluate these systems reside largely with their developers, and the science for rigorous audits does not yet exist. Governments can build capacity by leveraging their power as major, demanding customers, using procurement to support independent testing and shared benchmarks, provided the work is funded as research rather than expected on demand. Procurement law can then carry those emerging standards into the systems governments buy.

Capacity also includes the ability to inspect public procurement. An agency relying on a vendor's proprietary simulation to validate a system must be able to examine that environment, paired with real-world testing. In public benefits settings, the obligation to evaluate may also arise from due process values, rather than procurement alone. Much of this challenge extends across borders: The standards governing whether allied systems interoperate are being established now, and shaping them early is far more effective than reconciling incompatible regimes later. Ultimately, countries that build robust public ecosystems around simulation, robotics, and safety testing will outpace those focused exclusively on frontier model releases.

Conclusion

Spatial intelligence and world models are rapidly moving from research into the physical economy. The critical question is whether they mature within a narrow commercial and security logic or within a governance framework that channels their capabilities toward public value.

Handled responsibly, these models could strengthen economic resilience, expand scientific and industrial capacity, enhance national security, and support safer, more accountable embodied AI. Handled poorly, they could concentrate power over the simulation layer of the economy and the surveillance that attends it, exacerbate national security vulnerabilities to hostile actors, and accelerate unsafe deployment in high-stakes settings.

The difference comes down to choices made now: investing in shared infrastructure to prevent concentration, tying safeguards to deployment, and building the measurement science and public expertise needed to evaluate these systems. The central task is to shape the conditions under which this technology develops before its trajectory hardens.

Disclaimer

Stanford HAI has received financial support from several companies referenced in this brief, including Nvidia and Google. In addition, brief co-author Fei-Fei Li is currently on partial leave from Stanford University and serves as the co-founder and chief executive officer of World Labs.

Consistent with Stanford HAI's fundraising and research policies, no outside entity sets the institute's research agenda or determines the outcome or dissemination of its research. The authors conducted this research independently and retained full authority over the analysis, findings, recommendations, and editorial decisions presented in this brief.

Acknowledgment

The authors thank Alberto Tono for research assistance and Wenlong Huang for reviewing an early version of this brief.

Daniel Zhang is the chief of staff at the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

Russell Wald is the executive director at Stanford HAI.

Ehsan Adeli is an assistant professor (research) of psychiatry and behavioral sciences, of computer science (by courtesy), and of biomedical data science (by courtesy) at Stanford University, and a faculty affiliate of Stanford HAI.

Elena Cryst is the director of policy and society at Stanford HAI.

Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, professor of political science and of computer science (by courtesy), senior fellow and associate director at Stanford HAI, senior fellow at the Institute for Economic Policy Research, and director at the Regulation, Evaluation, and Governance Lab.

Caroline Meinhardt is the policy research manager at Stanford HAI.

Jiajun Wu is an assistant professor of computer science and of psychology (by courtesy) at Stanford University and a faculty affiliate at Stanford HAI.

Amy Zegart is the Morris Arnold and Nona Jean Cox Senior Fellow at the Hoover Institution, professor of political science (by courtesy), senior fellow at the Freeman Spogli Institute for International Studies, senior fellow and associate director at Stanford HAI, and founding director at the Hoover Institution's Tech Policy Accelerator.

Fei-Fei Li is the Sequoia Professor of Computer Science at Stanford University and the founding director and a senior fellow at Stanford HAI.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford, CA 94305-5008
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**