World Models Across Disciplines: A Functional Analysis from Ancient Precursors to Foundation AI

Fernando E. Rosas

April 23, 2026

Abstract

The term "world model" now spans a striking range of literatures, from model-based reinforcement learning and multimodal video generation to cognitive maps, internal models, mental models, and digital twins. This report argues that the apparent sprawl becomes intelligible once world models are treated functionally rather than by representational format alone. A world model is a structured surrogate of latent state, dynamics, and intervention-relevant regularities that enables a system to predict, explain, imagine, plan, control, or coordinate without relying exclusively on immediate interaction with the world itself. After situating the idea historically, from ancient reflections on mediated reality and imagination to Craik's "small-scale models" and modern AI, the report develops a minimal formal schema separating domain state, observations, actions, rewards, and model state. It then explains why world models matter technically under partial observability, delayed effects, costly experimentation, and long-horizon decision problems. On that basis, it compares three broad cross-disciplinary families—engineered world models in planning, control, robotics, model-based reinforcement learning, and digital twins; biological and cognitive world models in neuroscience, motor control, predictive processing, psychology, and prospection; and shared or applied world models in systems dynamics, management, and HCI—before turning to contemporary foundation world models trained on large-scale video, spatial, game, and robotic data. Across these traditions, the central dimensions of variation are explicit versus learned, internal versus external, and observation-oriented versus intervention-oriented modeling. The main conclusion is that "world model" does not name one privileged formalism. It names a recurrent architectural role: moving inference and counterfactual evaluation from direct world interaction into a manipulable surrogate space, with the strongest cases defined by action-sensitive competence rather than realism alone.

Contents

1 Introduction

The phrase "world model" now sits near the center of contemporary AI. It appears in model-based reinforcement learning, embodied agents, multimodal planning, robotics, video generation, and increasingly in discussions of general-purpose machine intelligence. Yet the term is unstable. In some papers it means a learned latent dynamics model; in others, a symbolic action theory, a cognitive map, a predictive generative model, a digital twin, or a user's conceptual understanding of a device. That instability is not merely terminological noise. It is evidence that multiple disciplines have repeatedly converged on a similar architectural solution to a similar class of problems.

The idea itself, however, is much older than the contemporary phrase. One useful modern anchor is Kenneth Craik's proposal that cognition depends on "small-scale models" of external reality [1]: internal surrogates that allow an organism to anticipate events by manipulating a model rather than waiting for the world itself to unfold. Even Craik is not the true beginning. Ancient philosophy repeatedly returned to precursor questions about mediated reality, imagination, dreaming, and representation. Plato's allegory of the cave asks what happens when agents mistake appearances for reality [2]. Zhuangzi's butterfly dream destabilizes the distinction between waking world and dreamed world [3]. Aristotle's discussions of *phantasia*, sleep, and dreams treat imagination as a

structured intermediary between sensation and thought [4, 5]. Sextus Empiricus turns such cases into a skeptical challenge: if waking and dreaming can come apart, then the world as experienced is not simply given but always mediated and revisable [6]. None of these traditions used the modern phrase "world model," but they clearly engaged problems about internal surrogates, simulation-like cognition, and the relation between appearance and reality.

Figure 1 places a few of these precursor moments on a schematic timeline. The continuity is functional rather than terminological: the recurring question is how representation, appearance, imagination, and reality relate.

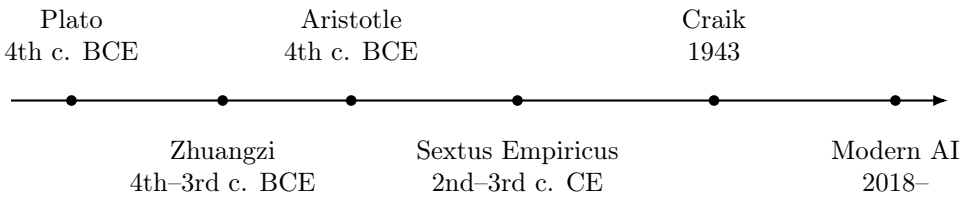

Figure 1: Schematic timeline of selected precursors to modern world-model language. The continuity is functional rather than terminological: recurring questions about mediated reality, imagination, and surrogate reasoning later reappear in scientific and engineering form.

These historical precursors matter for more than context. They show that the contemporary term did not emerge ex nihilo from machine learning; it expresses a much older intuition that intelligence improves when some of the burden of reasoning is shifted from the world itself into a manipulable surrogate. That need becomes especially acute under partial observability, delayed effects, branching futures, and costly or dangerous intervention. In such settings, purely reactive input-output mappings are often insufficient. Systems become more capable when they can update a model of hidden state, compare hypothetical futures, and evaluate actions before committing to them. The surrogate may be symbolic, dynamical, neural, linguistic, embodied, or shared across people and tools. The strongest cross-disciplinary commonality is therefore not prediction alone, but action-sensitive surrogate reasoning.

This report develops that claim in a deliberately cross-disciplinary way. It first offers a working definition of world models and a minimal formal schema separating domain state, observations, interventions, rewards, and model state. It then explains, in technical terms, why such models matter: they support belief updating under partial observability, off-line evaluation of hypothetical actions, sample-efficient policy improvement, abstraction for transfer, and causal explanation or coordination. On that basis, the report compares three broad historical families of work: engineered world models in planning, control, robotics, model-based reinforcement learning, and digital twins; biological and cognitive world models in neuroscience, motor control, predictive processing, psychology, and prospection; and shared or applied world models in systems dynamics, management, and HCI. It then turns to the contemporary foundation-model literature, where "world model" increasingly refers to large-scale predictive or generative systems trained on video, spatial, game, and robotic data.

The synthesis is grounded in representative research literatures spanning AI, robotics, neuroscience, cognitive science, and allied fields, and is supplemented only by a limited set of canonical references needed to make the conceptual bridges explicit. The goal is not bibliometric exhaustiveness. It is to clarify the main senses in which different communities have used world-model-like constructs, identify their formal commitments, and show where the underlying architectural idea is genuinely shared. The introduction situates the concept historically; Section 2 defines the term; Section 3 explains why it matters; Sections 4–6 survey the major cross-disciplinary families; Section 7 examines foundation-scale world models in contemporary AI; and Section 8 closes with a discussion that

compares the traditions and sharpens the paper's general conclusions.

2 What Are World Models?

Any attempt to define "world model" across disciplines immediately encounters a problem of heterogeneity. The literature does not point to one object so much as to a family of related constructs: symbolic action models in planning, dynamics models in control, learned latent simulators in model-based reinforcement learning, cognitive maps in neuroscience, internal models in motor control, generative models in predictive processing, mental models in psychology, and digital twins in engineering. These are not identical. They differ in ontology, learning mechanism, level of abstraction, and success criterion. Nevertheless, they are not unrelated either. They solve recognizably similar problems by providing structured surrogates that can be manipulated in place of the world itself.

2.1 A working definition

Working definition. A world model is a structured surrogate representation of the relevant state, dynamics, and causal or action-contingent regularities of an environment or task domain, such that an agent or system can use it off-line to predict, explain, imagine, plan, or control.

In reinforcement learning, especially in contemporary model-based RL, a "world model" usually means an action-conditioned predictive model of the environment that can be rolled forward in imagination. Concretely, the model encodes how latent state changes under actions, often together with reward and observation prediction, so that an agent can evaluate candidate trajectories without executing all of them in the real environment [19–21]. In this usage, the core question is not simply whether the agent has an internal representation of the world, but whether it has a model that is sufficiently accurate for planning, policy improvement, value estimation, exploration, or sample-efficient control. The emphasis is therefore operational and intervention-centered: a world model is useful because it lets the agent ask what would happen if it acted in a certain way.

In a broader representational sense, by contrast, "world model" refers to any structured surrogate of relevant worldly organization that supports cognition, whether or not the immediate task is reward maximization. On this reading, cognitive maps, internal models in motor control, predictive generative models, mental models, and digital twins all count as world-model-like constructs because they encode stable relations among hidden states, causes, events, actions, or constraints in a form that can guide inference and behavior. The focus here is less specifically on policy optimization and more on the general representational role: a world model is something through which a system interprets, anticipates, explains, or simulates its domain.

The working definition adopted here is deliberately broader than the RL usage, but narrower than "representation" in the most permissive sense. It does not require the model to be symbolic rather than neural, learned rather than hand-authored, internal rather than external, or comprehensive rather than task-specific. What it does require is that the representation preserve enough structure for the system to reason about unobserved, future, or hypothetical states without relying only on immediate experience. In that sense, the term "world" should not be taken too literally. In different communities the relevant domain may be a physical environment, a task space, a narrative situation, a social relation, a device interface, or an engineered process. The commonality lies in surrogate-enabled reasoning, not in a fixed choice of ontology.

2.2 Minimal formal structure

At a high level, most world-model settings are partially observed: the agent or system typically does not have direct access to the latent domain state w_t. Instead, it only has access to its own actions together with incoming observations and, in many decision-theoretic settings, rewards or costs. One can describe this in terms of an agent-specific *interface*: what matters operationally is not transparent access to the world's full underlying state, but the structured relation between action histories and the outcome histories available to that agent [18]. In our notation, we separate those available outcome channels into observations o_t and task-value signals r_t, while m_t denotes the internal or external model state maintained to bridge the gap between these interface variables and the hidden state w_t. Let a_t denote an action or intervention and let r_t denote a reward, cost, or utility signal associated with the consequences of action at time t. In reinforcement-learning settings, r_t is typically the scalar feedback that marks outcomes as more or less desirable and thereby defines what the agent is trying to maximize or avoid over time. Then a large class of world-model architectures can be written as

$$m_t \sim U_\theta(m_{t-1}, o_t, a_{t-1}),$$

where U_θ is the model update rule. This notation is intentionally permissive: in some systems m_t is a deterministic function of the previous model state, the new observation, and the preceding action, whereas in probabilistic state-space models or variational latent models it may be sampled, inferred, or otherwise stochastically generated conditional on those quantities. The model is useful insofar as it supports queries about hidden, future, or counterfactual structure. In predictive form this may look like

$$\hat{p}_\theta(w_{t+1:t+H}, o_{t+1:t+H}, r_{t:t+H-1} \mid m_t, a_{t:t+H-1}),$$

and in decision-theoretic form it may support a control or planning objective such as

$$\pi^* \in \arg\max_\pi \mathbb{E}_{\hat{p}_\theta,\pi}\left[\sum_{k=0}^{H-1} \gamma^k r_{t+k}\right],$$

corresponding to the sum of discounted future reward. This abstract template is intentionally permissive. In symbolic planning, m_t may simply be the explicit state description itself. In control, it may be a filtered state estimate. In model-based RL it may be a learned latent state. In predictive processing it may be a hierarchical belief state over hidden causes. In digital twins it may be an external, continuously updated estimate coupled to a real asset.

The importance of distinguishing w_t from m_t is that a world model need not reproduce the world one-for-one. The model only needs to preserve the aspects of structure relevant to the functions it serves. This is one lesson of MuZero-style planning [21], where the learned model need not reconstruct the raw observation stream, and one lesson of cognitive-map research [47, 48], where a representation can be useful because of its relational and predictive organization even if it is not a full sensory simulator.

Figure 2 summarizes this generic architecture. The critical separation is between the latent state of the domain and the model state maintained by the agent or artifact: the latter need not be a faithful copy of the former, but it must preserve enough structure to support useful queries.

2.3 What varies across communities

Different communities have taken this minimal structure and developed along various different axes.

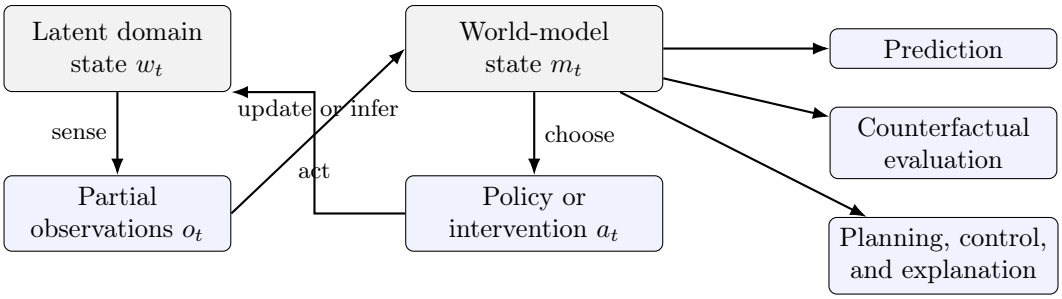

Figure 2: Generic architecture of a world model. The model state is updated from partial observation and used to answer off-line queries about hidden, future, or hypothetical structure.

1. *Given versus learned*. STRIPS and situation-calculus models [7, 8] make transition structure explicit and interpretable. Dreamer-style latent models [20] and related modern AI systems assume the agent needs to learn compact predictive structure from data.

2. *Internal versus external*. Cognitive maps, internal models, and mental models are ordinarily posited inside organisms or users, whereas digital twins [16] and robotic simulators are external artifacts.

3. *World-centric versus task-centric*. Some traditions aim to capture broad background structure, while others care only about preserving whatever supports successful planning or value estimation.

4. *Observational versus interventional orientation*. Predictive coding and perception-centered generative modeling are strongly concerned with explaining sensory input [54, 55, 57]. Classical planning, control, and model-based RL are more strongly organized around the question of what will happen under chosen actions [7, 10, 19].

5. *Individual versus shared*. Mental-model theory and motor-control theory usually concern individual agents [50, 60], whereas management science and systems dynamics emphasize causal models that can be shared, negotiated, and revised across groups [71, 72].

6. *Static versus dynamical* axis. Scripts, schemas, and situation models may primarily organize states and event relations [62, 65], whereas control theory and modern AI emphasize rollout, feedback, and transition dynamics.

2.4 What is not a world model

The preceding distinctions also clarify a negative boundary: not every representation that improves performance should be called a world model. A memory buffer, lookup table, embedding, classifier, retriever, or next-token predictor may all contain useful information about the world without yet constituting a world model in the stronger sense developed here. The issue is not whether the representation stores regularities, but whether it supports structured surrogate reasoning about hidden state, possible futures, and the effects of interventions.

One helpful diagnostic is to ask whether the representation supports three things.

1. *Latent organization.* Does it encode more than surface associations between observed inputs and outputs? A world model ordinarily compresses some hidden or relational structure of the domain.

2. *Intervention-sensitive querying.* Can it be queried under hypothetical actions, goals, perturbations, or causal assumptions, rather than only under the passive continuation of the current input stream?

3. *Multi-step surrogate reasoning.* Can it be used to project, compare, explain, or evaluate extended trajectories, rather than merely produce the next local prediction?

A narrow predictor of the form

$$\hat{y}_{t+1} = f_\theta(o_{\leq t})$$

may be statistically useful while still failing all three tests. It may forecast the next observation without representing enough latent organization to support intervention, explanation, or long-horizon imagination. In other words, it can exploit correlations in the observation stream without providing a manipulable surrogate of the domain itself.

By contrast, a world model supports queries parameterized by hypothetical actions, goals, or latent causes, for example

$$Q_\theta(m_t, a_{t:t+H-1}, g) \mapsto (\hat{\tau}_{t:t+H}, \hat{o}_{t:t+H}, \hat{r}_{t:t+H-1}),$$

where g denotes an optional goal specification and the output is an anticipated latent trajectory $\hat{\tau}$, together with predicted observations and task-relevant consequences under a contemplated course of action. The query interface need not be symbolic or hand-authored. In classical planning it may appear as explicit successor-state computation; in model-based RL as latent imagination; in a digital twin as simulated stress testing under alternative operating conditions.

This functional criterion also clarifies important boundary cases. A sequence model, video generator, or large language model is not automatically a world model merely because it predicts coherent continuations. But such a system can become world-model-like if it learns stable latent state, supports action-conditioned or counterfactual rollout, and preserves enough structure for downstream planning, explanation, or control. Similarly, a database, memory system, or embedding space may be a component of a world model without itself being one. The term is therefore best reserved for representations that do more than store correlations: they must preserve enough relational or dynamical organization that a system can reason *with* them rather than merely react *through* them.

3 Why World Models Matter

World models are important for agents because real environments and task domains are almost never fully transparent to agents acting in them. Consequences are delayed, states are only partially observed, action spaces branch combinatorially, and trial-and-error in the real world is often expensive, dangerous, or irreversible. Under those conditions, intelligence depends on the ability to carry some of the burden of reasoning out of the world and into a surrogate model. That claim can be made precise in several complementary ways.

Figure 3 provides a schematic overview of the logic developed in this section: difficult decision conditions create pressure for model-based operations, and those operations open capabilities that would otherwise be unavailable or too costly.

3.1 Prediction, hidden state, and partial observability

The first reason world models matter is that intelligent action is ordinarily a problem of inference under partial observability. If the relevant state of the environment is hidden, then current observation

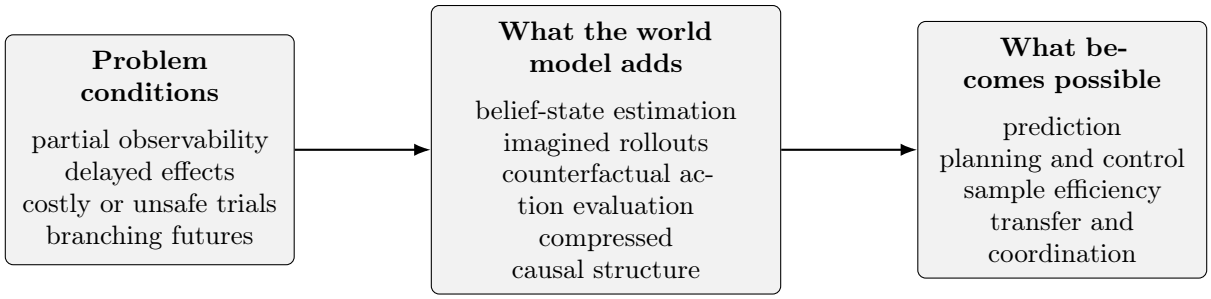

Figure 3: Why world models matter. They convert difficult environmental conditions into tractable internal or external computations, which in turn support prediction, planning, and coordination.

alone is insufficient for optimal prediction or control. In a partially observed controlled system, one can write the belief-state update as

$$b_{t+1}(w') \propto O(o_{t+1} \mid w') \sum_{w} T(w' \mid w, a_t)\, b_t(w),$$

where b_t is a distribution over latent states, T is the transition model, and O is the observation model. Even when an agent does not literally implement Bayesian filtering, the structural point remains: some internal or external representation must integrate past observations and action history into a predictive state.

This is exactly the role played by forward and internal models in sensorimotor control [50–53], by predictive generative models in predictive coding and active inference [54–57], and by latent-state models in contemporary AI [20]. Without some such model, the system is restricted to myopic stimulus-response mappings. With one, it can estimate hidden causes, filter noise, and anticipate what is likely to happen next.

3.2 Counterfactual reasoning and planning

The second reason world models matter is that they support counterfactual reasoning: what would happen if I did this rather than that? In formal planning terms, the problem is to evaluate action sequences without physically enacting all of them. Given a model $\hat{T}$ and reward or cost model $\hat{R}$, one can define a planning objective

$$(a_t^*, \ldots, a_{t+H-1}^*) \in \arg\max_{a_{t:t+H-1}} \mathbb{E}_{\hat{T},\hat{R}} \left[\sum_{k=0}^{H-1} \gamma^k \hat{r}_{t+k} \right].$$

The model's value is precisely that it lets the agent compare hypothetical futures in model space.

This logic unifies several traditions that otherwise look very different. In classical AI it appears as search over symbolic transition systems [7, 8]. In model predictive control it appears as receding-horizon optimization over learned or engineered dynamics. In Dyna-style and Dreamer-style systems [10, 20], imagined experience generated by the model directly improves the policy. In MuZero [21], the world model is useful because it preserves the structure needed for search and value estimation, not because it reconstructs all sensory detail. In each case, the crucial gain is the same: the system can deliberate over possible futures before committing to one.

Psychology and cognitive science reach a related point from another direction. Mental models are valuable because they let reasoners test possibilities, evaluate counterfactuals, and determine whether conclusions hold across represented cases [60, 61]. The functional commonality with planning is

strong even when the representation is linguistic, diagrammatic, or possibility-based rather than control-theoretic.

3.3 Control, sample efficiency, and safe action

World models also matter because they decouple improvement in behavior from irreversible interaction with the world. Let $\mathcal{D}_{\text{real}}$ denote a set of real transitions collected from the environment and let $\mathcal{D}_{\text{model}}$ denote imagined transitions generated by a learned or engineered model:

$$\mathcal{D}_{\text{real}} = \{(o_t, a_t, o_{t+1}, r_t)\}_{t=1}^{N_{\text{real}}}, \qquad \mathcal{D}_{\text{model}} \sim \hat{p}_\theta(\cdot \mid \mathcal{D}_{\text{real}}).$$

Once a model is available, policy improvement can draw on both datasets rather than only the real one. This is the basic logic behind model-based sample efficiency: the number of policy updates or planning evaluations need not equal the number of costly physical interactions.

The engineering significance is especially clear when real interaction cost dominates model-evaluation cost:

$$C_{\text{total}} = N_{\text{real}}\, c_{\text{real}} + N_{\text{model}}\, c_{\text{model}}, \qquad c_{\text{model}} \ll c_{\text{real}}.$$

In robotics and control, where c_{real} may include wear, safety risk, or time, the asymmetry can be decisive [12–15]. In digital twins it appears as stress testing and intervention screening before deployment [16]. In motor neuroscience it appears as the biological capacity to compensate online for delay and noise through predictive internal models [50]. Across these cases, world models matter because they move a large portion of optimization and error discovery into surrogate space.

3.4 Abstraction, transfer, and generalization

Another reason world models matter is that they provide a basis for abstraction. Raw observations are often too high-dimensional, too local, or too entangled to support transfer. A good world model compresses experience into a representation that preserves future-relevant regularities. In general form, one seeks a representation map ϕ such that

$$P(o_{t+1:t+H}, r_{t:t+H-1} \mid w_t, a_{t:t+H-1}) \approx P(o_{t+1:t+H}, r_{t:t+H-1} \mid \phi(w_t), a_{t:t+H-1}).$$

for the horizons and interventions that matter to the task. This is why the best world models are often not exhaustive replicas of reality. They are useful abstractions.

The point is visible in several literatures. Cognitive-map research shows that relational structure can support shortcutting, transfer, and generalization beyond memorized trajectories [43, 47, 48]. Mental-model theory shows that structured possibility representations support inference beyond surface similarity [60]. MuZero-style results show that decision-relevant abstraction can be enough for strong planning [21]. Modern AI world models pursue precisely this compression problem in learned latent spaces [19, 20]. What matters is not full duplication of the world but preservation of the structure that supports successful reuse under novel goals and contexts.

3.5 Explanation, communication, and coordination

World models matter not only for solitary prediction and control, but also for explanation and coordination. A model that represents causal or mechanistic structure allows a system to answer why-questions, diagnose failures, and communicate assumptions. In causal notation, one might write a world model as a collection of structural equations

$$x_i = f_i(\text{pa}(x_i), u_i),$$

where each variable depends on its parents and exogenous disturbances. Whether or not a community uses this exact formalism, the underlying idea recurs: the model exposes dependencies that can be inspected, discussed, and manipulated.

This explanatory and coordinating role is central in several traditions. Mental-model theory and situation-model research explain how humans reason and maintain coherence by operating over structured internal states [60, 65]. In HCI, useful user mental models enable prediction, transfer, and error recovery even when the user does not possess a complete technical specification [69, 70]. In systems dynamics and organizational learning, shared mental models matter because intervention quality depends on how stakeholders represent feedback loops, delays, and causal structure [71, 72]. These cases show that world models are not just computational tools for agents. They are also mediating objects for interpretation, alignment, and collective action.

3.6 Why the concept matters scientifically

The broader scientific importance of world models is that they offer a unifying lens on intelligence across artificial and natural systems. Once the concept is stated functionally, a large number of otherwise separate research programs can be seen as variations on one architectural idea: successful behavior under uncertainty depends on structured surrogate representations of latent organization and change. The details differ. Some models are symbolic, some neural; some internal, some external; some individual, some shared. But the recurrent function is stable. World models matter because they explain how systems break out of purely reactive behavior and acquire anticipatory, counterfactual, and generalizing competence.

4 Engineered World Models

The first major family comprises engineered world models: representations built as artifacts for planning, control, search, monitoring, or policy optimization. This family includes symbolic action models, dynamics models, learned latent simulators, and externalized twins. What unifies them is that they are constructed to support intervention in artificial or engineered environments. Formally, the recurring pattern is a model of state evolution,

$$s_{t+1} \sim T_\theta(s_t, a_t),$$

possibly augmented with observations o_t, rewards r_t, constraints, or latent variables. The technical differences across subfields concern what counts as state, how T_θ is represented, and how the model is optimized or exploited.

Figure 4 gives a structural map of the engineered family surveyed in this report.

4.1 Symbolic Models of Action and Commonsense Worlds

4.1.1 Planning and action

In classical AI, the closest ancestor of the modern world model is the explicit symbolic model of states, actions, and consequences. The world is represented propositionally or logically; actions change the state; and planning is the search for action sequences that transform an initial state into a goal state. This is the tradition of the situation calculus and STRIPS [7, 8].

The canonical abstraction is a transition system $(\mathcal{S}, \mathcal{A}, T, G)$, with symbolic states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, transition function T, and goal condition G. In STRIPS-like planning, an action is represented as

$$a = (\mathrm{Pre}(a), \mathrm{Add}(a), \mathrm{Del}(a)),$$

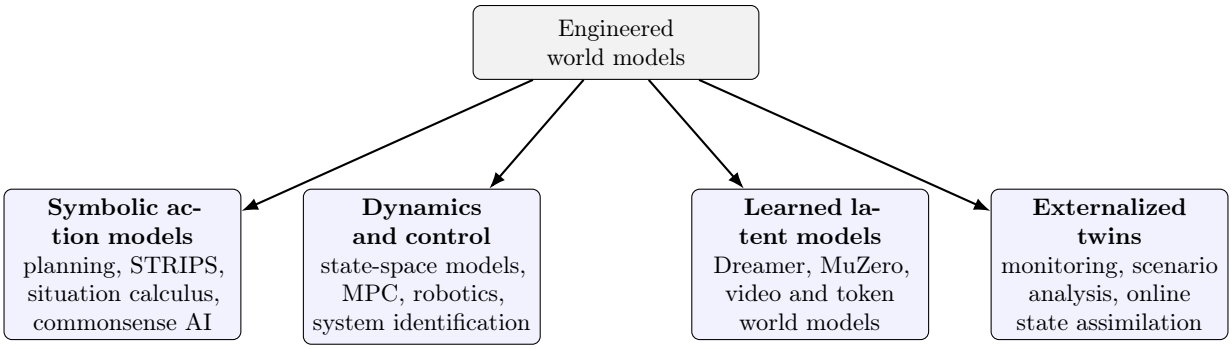

Figure 4: Main branches of engineered world-model research. The family spans explicit symbolic action systems, continuous dynamics models, learned latent simulators, and externally deployed digital twins.

and when $\mathrm{Pre}(a) \subseteq s_t$ the successor state is

$$s_{t+1} = (s_t \setminus \mathrm{Del}(a)) \cup \mathrm{Add}(a).$$

The optimization problem is then to find an action sequence $(a_0, \ldots, a_{H-1})$ such that $G \subseteq s_H$ while minimizing some cost $\sum_{t=0}^{H-1} c(a_t)$.

Here the model is explicit, compositional, and intervention-centric. The central question is not merely "what observations are likely next?" but "what changes if I do this?" The model succeeds when it supports sound or useful search over possible state trajectories. This classical planning sense remains highly relevant because many contemporary LLM-agent systems have rediscovered it: once agents need reliable long-horizon behavior, developers often reintroduce explicit preconditions, effects, simulators, or search spaces.

Sutton's Dyna architecture [10] is especially important as a bridge. It preserved the action-transition view of the world while combining it with learning, thereby prefiguring much of later model-based reinforcement learning.

4.1.2 Commonsense reasoning and naive physics

Closely related, but conceptually distinct, is the commonsense AI tradition. In this line of work, the model is not only a task-specific action system but a broader representation of the everyday physical and social world: persistence, containment, support, motion, time, and ordinary causation. Hayes' *Naive Physics Manifesto* is canonical here [9]. The goal is not simply to solve a benchmark planning instance but to encode enough qualitative world structure for robust inference across ordinary situations.

Technically, this tradition extends symbolic action models by enriching the ontology of states and transitions. Instead of a narrow action calculus, one seeks a background theory over fluents $F_i(s)$, persistence assumptions, and qualitative relations such as *inside*, *supports*, or *before*. In situation-calculus notation, the problem is to reason about statements like

$$F(\mathrm{do}(a, s))$$

without having to restate every unaffected fact after every action. This is where issues like the frame problem become central: a world model must specify what changes, but also what stays the same.

This tradition deserves inclusion in a world-model survey because it articulates one of the strongest symbolic versions of the same aspiration now expressed in learned-model language: intelligence

requires structured background knowledge about how the world behaves. Modern commonsense surveys, such as Davis and Marcus [11], make clear that this problem remains open.

4.1.3 Consequences and implications

The practical consequence of symbolic world models is systematic intervention over combinatorial spaces. These models enabled researchers to synthesize plans, verify action sequences, reason explicitly about preconditions and side effects, and state counterfactuals in a form that can be checked rather than merely intuited. Commonsense extensions then broadened the scope from narrow planning domains to more general physical and causal reasoning. In effect, these models made it possible to ask not just "what should I do?" but "under what structural assumptions would doing this work at all?" That shift remains foundational for verification, symbolic planning, tool use, and neurosymbolic agent design.

4.2 Dynamics Models in Control, Robotics, and Engineering

4.2.1 Forward dynamics, system identification, and MPC

In control and robotics, a world model is typically a dynamics model: a plant model, forward model, state-space system, or simulator that predicts the effects of control inputs. This tradition includes system identification, model predictive control, trajectory optimization, adaptive control, and data-driven dynamics learning.

The standard form is a controlled dynamical system

$$x_{t+1} = f_\theta(x_t, u_t) + w_t, \qquad y_t = h_\theta(x_t) + v_t,$$

where x_t is the latent system state, u_t the control input, y_t the observation, and w_t, v_t process and observation noise. In model predictive control one solves

$$\min_{u_{0:H-1}} \sum_{t=0}^{H-1} \ell(x_t, u_t) + \ell_f(x_H)$$

subject to the model dynamics and any state or control constraints. The robotics literature shows that this lineage is not peripheral but foundational. It includes control foundations from Kalman-style filtering through predictive control, trajectory optimization methods such as iterative LQG and differential dynamic programming, and later data-driven approaches such as Koopman-based models and neural MPC.

In this community, models are evaluated by control utility: rollout fidelity matters because it determines whether planning in model space transfers back to the physical system. This sense overlaps with classical AI in its emphasis on intervention, but it differs in representation and scale. Classical planning usually assumes discrete symbolic states and exact or hand-authored transitions. Control and robotics typically deal with continuous states, uncertainty, partial observability, real-time constraints, and approximate models.

4.2.2 Robot world models

Modern robot-learning papers increasingly adopt the literal phrase "world model" for learned models of visual and sensorimotor dynamics. This literature includes visual-foresight systems [13, 14], learned real-world robot models such as DayDreamer [15], and a growing family of zero-shot, pre-trained, or

interactive robot simulators. These works use world models to make real robots more data-efficient, safer, and better able to plan before committing to physical action.

Technically, robot world models often introduce a latent state z_t inferred from high-dimensional observations:

$$z_t \sim q_\phi(z_t \mid o_{\le t}, a_{<t}), \qquad z_{t+1} \sim p_\theta(z_{t+1} \mid z_t, a_t).$$

A planner or policy is then optimized with respect to imagined rollouts in latent space:

$$\max_\pi \ \mathbb{E}_{p_\theta,\pi}\left[\sum_{t=0}^{H-1} \gamma^t \hat{r}_\theta(z_t, a_t)\right].$$

This robotics literature is where older control-theoretic modeling and newer learned latent-modeling approaches visibly converge. The model is no longer only a conventional system-identification object, but neither is it merely a general generative model. It is an action-conditioned predictive surrogate whose value is judged by downstream behavior.

4.2.3 Digital twins as externalized world models

Engineering introduces a further extension: the digital twin. A digital twin is an external model of a real system, often updated with live data and used for monitoring, forecasting, scenario analysis, and intervention testing [16]. Unlike cognitive or neural world models, digital twins are not typically internal representations within an agent. They are deployed artifacts.

Mathematically, a digital twin can be understood as a model-plus-assimilation loop. If $\hat{x}_t$ denotes the twin's estimate of the physical state, a simple update has the form

$$\hat{x}_{t+1} = f_\theta(\hat{x}_t, u_t) + K_t\big(y_t - h_\theta(\hat{x}_t)\big),$$

where the second term corrects the simulated state using incoming observations. In that sense a twin is not just a simulator but an online estimator coupled to the real system.

Conceptually, digital twins belong in this family because they are counterfactual-capable surrogates for portions of the world. They support exactly the kind of off-line-if-needed, in-the-loop-if-possible reasoning that motivates world models elsewhere. The recent AI literature increasingly treats digital twins and learned world models as adjacent or even convergent technologies, especially in infrastructure, mobile networks, and embodied systems.

4.2.4 Cognitive and developmental robotics

The survey by Taniguchi and colleagues [17] is important because it explicitly connects robotics world models to predictive coding, active inference, and developmental cognition. This bridge clarifies that robotics has not only an engineering lineage but also a cognitive one. In developmental robotics, a world model is often conceived as the predictive core of an autonomous learner that acquires increasingly structured expectations about action, perception, and social interaction.

One useful formalization here is the predictive-processing view of robot learning, in which the robot minimizes a prediction or free-energy objective over latent causes:

$$F[q] = \mathbb{E}_{q(z)}[\log q(z) - \log p_\theta(o, z)].$$

The same system then uses action to reduce expected mismatch between predicted and observed sensorimotor trajectories. This makes developmental robotics a genuine bridge between control-theoretic prediction and cognitive theories of generative modeling.

4.2.5 Consequences and implications

The central consequence of dynamics-based world models is the ability to move expensive, dangerous, or irreversible trial-and-error into model space. These models made it possible to do receding-horizon optimization, safe trajectory planning, online adaptation, counterfactual testing on twins, and robot learning with dramatically fewer physical interactions than model-free alternatives typically require. More broadly, they shifted engineering from reactive control toward anticipatory control, where candidate interventions can be compared under model rollouts before being deployed on the real system.

4.3 Learned World Models in Model-Based Reinforcement Learning and Modern AI

In contemporary AI, "world model" usually refers to a learned model of environment dynamics that can be rolled forward in imagination. The model may operate in pixel space, latent state space, object-centric state space, token space, or some multimodal combination thereof. The core use cases are policy improvement, value estimation, planning, exploration, and sample efficiency.

4.3.1 From classic model-based RL to latent imagination

The AI and robotics literature makes clear that the current usage sits on top of a longer model-based RL lineage: stochastic value gradients, value prediction networks, probabilistic dynamics models, model-based policy optimization, and related work. Ha and Schmidhuber's *World Models* paper [19] is therefore both a milestone and a naming event: it made the term memorable and tied it to unsupervised representation learning plus imagined control.

In a generic latent world-model setup, one learns

$$z_{t+1} \sim p_\theta(z_{t+1} \mid z_t, a_t), \qquad \hat{r}_t = r_\theta(z_t, a_t), \qquad \hat{o}_t \sim p_\theta(o_t \mid z_t),$$

and then optimizes a policy against the learned model:

$$J(\pi) = \mathbb{E}_{p_\theta, \pi}\left[\sum_{t=0}^{H-1} \gamma^t \hat{r}_t\right].$$

Dreamer [20] extended this idea by showing that long-horizon behavior can be learned effectively from latent imagination alone. MuZero [21] then sharpened the point that a useful world model need not reconstruct the raw world. It need only preserve the task-relevant aspects of dynamics required for planning and value estimation.

4.3.2 Scaling out: transformers, video, language, and multimodal agents

Recent work broadens the term further. Recent AI research shows growth in transformer-based sequence world models, video-generative or video-predictive world models, object-centric and geometry-aware models, multimodal world models for embodied AI, and world models coupled to language models or code generation.

One can write the sequence-model view as an autoregressive factorization,

$$p_\theta(\tau) = \prod_{t=1}^{T} p_\theta(x_t, a_t \mid x_{<t}, a_{<t}),$$

or in latent-token form

$$p_\theta(z_{1:T}, a_{1:T}) = \prod_{t=1}^{T} p_\theta(z_t, a_t \mid z_{<t}, a_{<t}).$$

Examples include language-grounded dynamics learning [22], LLM-generated world-model programs [23], and LLM-assisted planning with explicit learned world models [24]. These extensions enlarge the domain from classic RL environments to web interfaces, navigation, code execution, manipulation, and multimodal instruction following.

4.3.3 What changed in the modern AI usage

Relative to older traditions, the modern AI usage shifts emphasis in three ways. First, the models are mostly learned rather than hand-authored. Second, latent representation quality is judged by downstream decision utility rather than by interpretability or symbolic fidelity. Third, there is an increasing tendency to treat generalist perceptual and generative models as candidate world models, especially when they can support action-conditioned rollout or structured reasoning.

This shift can be expressed as a move from exact transition fidelity to sufficient predictive structure. If the true environment dynamics are P and the learned model is $\hat{P}_\theta$, then the modern question is often not whether $\hat{P}_\theta = P$, but whether for the relevant decision functional $\mathcal{J}$,

$$\mathcal{J}(\pi; \hat{P}_\theta) \approx \mathcal{J}(\pi; P)$$

for the policies and horizons of interest. This is also why evaluation has become a central issue. Recent benchmark and evaluation work reflects the community's recognition that plausible-looking prediction is not enough. A world model for action must preserve task-relevant causal structure, memory consistency, and control affordances.

4.3.4 Consequences and implications

Learned world models in AI made several things newly feasible at scale: policy improvement through imagined rollouts rather than only direct experience; planning in latent spaces too compact for raw-state search; transfer across tasks through reusable predictive structure; and multimodal reasoning in environments where symbolic hand-modeling would be intractable. They also enabled a new research program around generalist agents: instead of training a separate controller for each setting, one can ask whether a single learned dynamics prior can serve planning, control, language grounding, and evaluation across many domains.

Section Summary: Engineered World Models

Main definitions. Engineered world models are artifact-level representations of state, dynamics, constraints, and action consequences, built to support planning, control, search, monitoring, or policy optimization.

Main features.

- Explicit symbolic transitions in planning and commonsense AI.
- Continuous state-space dynamics in control and robotics.
- Learned latent transition models in model-based RL and multimodal AI.
- Externalized online surrogates in digital twins.

Main achievements.

- Automated plan synthesis and formal reasoning about action.
- Receding-horizon control, trajectory optimization, and safer robot learning.
- Sample-efficient policy learning through imagination and latent planning.
- Counterfactual monitoring and intervention testing through digital twins.

5 Biological and Cognitive World Models

The second major family comprises biological and cognitive world models: representations posited to explain how organisms perceive, predict, remember, reason, imagine, and act. In this family the models are typically internal rather than external, and the emphasis shifts from engineered performance to neural, behavioral, or cognitive function. The formal objects vary widely, but they repeatedly involve latent structure that compresses experience while preserving future-relevant regularities.

Figure 5 summarizes the main subfamilies discussed in the next two sections.

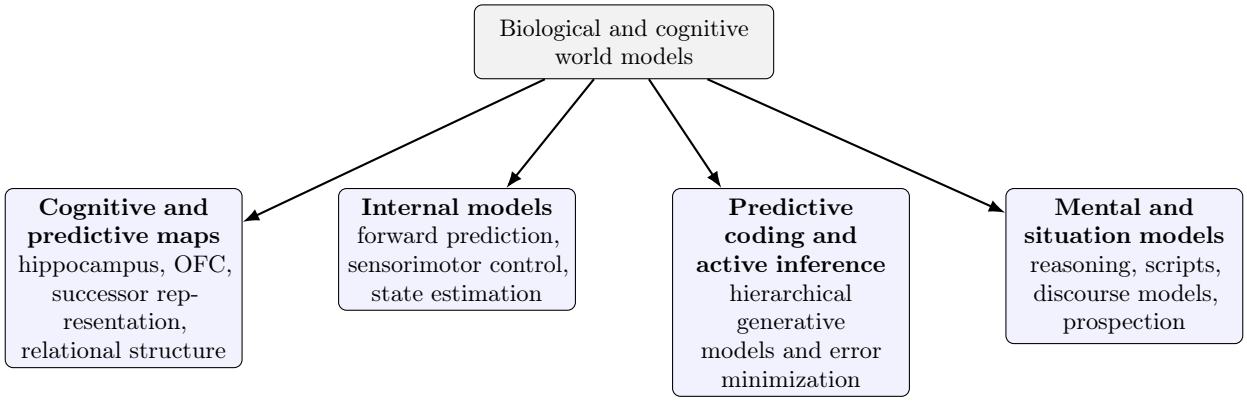

Figure 5: Main branches of biological and cognitive world-model research. These traditions are mostly internalist, but they vary sharply in formalization, neural scope, and behavioral role.

5.1 Neuroscience

The neuroscience literature shows that no single neuroscientific construct corresponds exactly to the modern AI phrase "world model." Instead, several partially overlapping families do the relevant work.

5.1.1 Cognitive maps and predictive maps

The most visible family is the cognitive map tradition. Beginning with Tolman's cognitive maps [43] and O'Keefe and Dostrovsky's hippocampal place-map evidence [44], this line of research investigates how animals and humans represent the latent structure of environments in ways that support navigation and flexible inference.

Over time, the map concept expanded beyond literal space. Orbitofrontal-cortex work argued that task structure can be represented as a cognitive map [45, 46]. Behrens et al. [48] framed cognitive maps as general-purpose organizing structures for flexible behavior. Stachenfeld et al. [47] gave a predictive reinterpretation through the successor representation,

$$M^{\pi}(s, s') = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} \mathbf{1}\{s_t = s'\} \,\middle|\, s_0 = s\right],$$

which then yields values through

$$V^{\pi}(s) = \sum_{s'} M^{\pi}(s, s')\, r(s').$$

The Tolman-Eichenbaum Machine [49] then proposed a unifying computational account of space and relational memory.

In this family, the model is often less a high-fidelity simulator than a structured state space with predictive consequences. What matters is that it supports transfer, generalization, planning, or compositional inference.

5.1.2 Internal models in sensorimotor control

A second family is the internal-model tradition in motor control. Wolpert, Ghahramani, Jordan, Miall, and Kawato argued that the brain uses forward and inverse models to predict the sensory consequences of action and to stabilize control under delay and noise [50–53]. Here the model is tightly coupled to action and state estimation.

The standard control-theoretic picture is that the nervous system maintains an internal estimate $\hat{x}_t$ and predicts sensory consequences via

$$\hat{x}_{t+1} = f(\hat{x}_t, u_t), \qquad \hat{y}_{t+1} = h(\hat{x}_{t+1}).$$

Prediction errors $y_{t+1} - \hat{y}_{t+1}$ can then be used for correction. Compared with cognitive maps, this is a narrower but more operational sense of world modeling. The model succeeds when it helps produce accurate, robust movement. Compared with predictive processing, it is also more localized and more directly linked to motor computation.

5.1.3 Predictive coding and active inference

A third family is predictive coding and active inference. Rao and Ballard's predictive-coding account [54] and Friston's free-energy framework [55–57] treat the brain as maintaining a hierarchical generative model of hidden causes and continuously updating that model through prediction error.

In predictive-coding form, one often writes layerwise prediction errors such as

$$\epsilon_{\ell} = \mu_{\ell} - f_{\ell}(\mu_{\ell+1}),$$

with neural dynamics that reduce a weighted sum of squared errors. In free-energy form, the variational objective is

$$F[q] = \mathbb{E}_{q(s)}[\log q(s) - \log p(o, s)],$$

and perception corresponds to approximate Bayesian inference while action can be cast as reducing expected future free energy. In this perspective, the term most often used is "generative model" rather than "world model." But functionally the overlap is obvious: the model explains sensory input, anticipates future input, and in active-inference formulations also shapes action. Clark's review [58] is especially useful for situating this framework within broader cognitive science.

5.1.4 Replay, successor representation, and model-based control

The neuroscience corpus also shows that replay, planning, successor representations, and model-based control now form a dense intermediate territory between maps and generative models. Replay studies ask how future trajectories are internally reactivated. Successor-representation work provides a predictive basis for map-like state representations. Orbitofrontal-hippocampal research examines task-state coding and flexible planning.

Technically, replay can be viewed as the internal generation of trajectories

$$\tilde{\tau} = (\tilde{s}_0, \tilde{s}_1, \ldots, \tilde{s}_H)$$

from a learned transition structure, either for policy evaluation, credit assignment, or prospective search. Together, these lines suggest that biological world modeling is distributed across multiple specialized but coordinated mechanisms rather than implemented as one monolithic module.

5.1.5 Consequences and implications

The neuroscientific payoff of these models has been substantial. Cognitive-map and predictive-map theories made it possible to explain flexible navigation, relational inference, and generalization beyond simple stimulus-response learning. Internal-model theories explained how biological agents achieve fast, stable sensorimotor control despite delays and noise. Predictive coding and active inference reframed perception as inference rather than passive registration. Replay and successor-representation accounts, meanwhile, provided concrete mechanisms for prospective planning, offline learning, and transfer across tasks and spaces, including non-spatial ones.

5.2 Cognitive Science and Psychology

5.2.1 Mental models

In cognitive psychology, mental-model theory treats reasoning as the construction and manipulation of internal representations of possible situations [59, 60]. These models support deduction, causal inference, counterfactual reasoning, and explanation. They are often not full environment simulators. Instead, they are structured representations of the possibilities relevant to a task.

One compact formal reading is that a reasoner maintains a finite set of possibilities

$$M = \{\omega_1, \ldots, \omega_k\},$$

and evaluates an inference by checking whether the target conclusion φ holds in every possibility currently represented:

$$M \models \varphi \quad \text{iff} \quad \forall \omega \in M,\ \omega \models \varphi.$$

This sense is narrower than some uses of the term in AI, but it shares the key architectural commitment: intelligent behavior depends on internal structure-preserving surrogates rather than on surface associations alone. Khemlani et al. [61] extend this to causal reasoning directly.

5.2.2 Schemas, scripts, and situation models

Another major tradition covers schemas, scripts, and situation models. Schank and Abelson's scripts [62] represent stereotyped event sequences. Abelson's later discussion [63] clarifies their psychological status. Situation-model research in discourse comprehension argues that readers build dynamic internal representations of described events and update them as narratives unfold [64, 65].

One useful abstraction is a recursively updated discourse state:

$$m_t = U(m_{t-1}, e_t),$$

where m_t is the current situation model and e_t is the incoming event, sentence, or proposition. Scripts and schemas then act as priors over expected event transitions,

$$p(e_{t+1} \mid m_t, \sigma),$$

with σ denoting the active schema or script. This family is world modeling for comprehension rather than control. The model tracks who is where, what has happened, what is likely next, and what remains coherent with prior information. In modern terms, it is an event-structured predictive model of a narrative world.

5.2.3 Embodied simulation and prospection

Barsalou's perceptual symbol systems [66] and the literature on episodic future thinking and scene construction [67, 68] provide yet another sense. Here cognition relies on the partial re-enactment or reconstruction of perceptual and episodic structure. The system does not just infer abstract propositions; it constructs scenes, possible futures, and imagined episodes.

Although this literature is less often expressed with compact equations, its computational commitment is clear: future-oriented cognition works by recombining stored traces into candidate scenes with task-relevant detail. A generic formulation is a constructive generative process,

$$z_{\text{future}} \sim p(z \mid z_{\text{memory}}, g, c),$$

where a future scene representation is generated from memory traces, current goals g, and contextual constraints c. This strand is especially relevant to current AI discussions because it highlights imagination as a central use case for world models. RL researchers often describe planning in terms of imagined rollouts. Psychology and memory research long ago described a closely related phenomenon as episodic simulation.

5.2.4 Consequences and implications

The major consequence of cognitive and psychological world-model theories is explanatory leverage over forms of intelligence that are not well described by reactive association alone. Mental-model theory helped explain structured reasoning and counterfactual thought. Situation-model theory explained predictive discourse comprehension and coherence maintenance. Script and schema theories explained how prior world knowledge constrains interpretation and memory. Embodied simulation and prospection research clarified how agents project themselves into novel futures rather than

merely extrapolating symbols. Together these traditions made it possible to treat imagination, understanding, and reasoning as operations over structured internal surrogates rather than as isolated stimulus-response mappings.

Section Summary: Biological and Cognitive World Models

Main definitions. Biological and cognitive world models are internal representations posited to support perception, action, memory, reasoning, or imagination in organisms and human thinkers.

Main features.

- Relational or predictive state spaces in cognitive-map research.
- Forward and inverse models in sensorimotor control.
- Hierarchical generative models in predictive coding and active inference.
- Possibility-based, script-based, and scene-based representations in psychology.

Main achievements.

- Explaining navigation, replay, and transfer in latent task spaces.
- Explaining fast motor control under uncertainty and delay.
- Recasting perception as inference under generative models.
- Explaining reasoning, comprehension, and prospection as structured simulation.

6 Shared and Applied World Models

The third major family comprises shared and applied world models. These are neither purely internal cognitive structures nor purely technical simulators. Instead, they are models used by groups, institutions, designers, or end users to coordinate action, interpret systems, and make interventions under uncertainty.

Figure 6 highlights the loop characteristic of this family: world models mediate between systems, stakeholders, observations, and collective interventions.

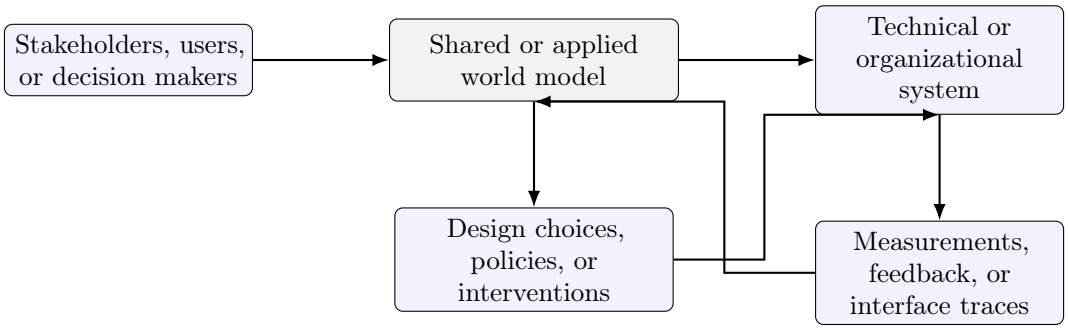

Figure 6: Shared and applied world models as coordination devices. The model does not merely predict a system; it helps align stakeholders, interventions, and feedback.

6.1 Systems Dynamics, Management, and Shared Mental Models

The systems-dynamics and organizational-learning literature adds one more important layer. Here the relevant object is the decision-maker's causal picture of a dynamic system with stocks, flows, delays, and feedback loops. Senge's *The Fifth Discipline* [71] popularized "mental models" as one of

the central obstacles to and resources for organizational learning. Doyle and Ford [72] attempted to formalize the concept for dynamic systems research.

Formally, this tradition often represents systems as stock-flow models:

$$\dot{x}(t) = S\, v(x(t), u(t), t),$$

or in discrete time,

$$x_{t+1} = x_t + \Delta t\big(f^+(x_t, u_t) - f^-(x_t, u_t)\big),$$

where x_t denotes stocks, f^+ inflows, and f^- outflows. Shared mental models can then be viewed as partial causal graphs over these variables, with agents differing in which nodes, delays, and feedback loops they include. This tradition matters because it introduces the individual-versus-shared dimension in a particularly clear form. World models need not be private cognitive structures. They can be shared, negotiated, externalized, and revised through group model building. In that respect, this literature connects cognitive notions of modeling with engineering and policy practice.

6.1.1 Consequences and implications

The payoff of shared dynamic world models is improved coordination under complexity. They enabled researchers and practitioners to reason about delayed effects, feedback loops, policy resistance, and emergent behavior that are hard to grasp locally. In practical terms, they made it possible to diagnose why interventions fail, compare competing causal pictures across stakeholders, and design policies that account for system-wide dynamics rather than only immediate local effects.

6.2 HCI and conceptual understanding

HCI uses "mental model" in a more applied way. Norman [69] and Rouse and Morris [70] focus on the user's explanatory understanding of how a device or interface works. Here the question is not whether the model is neurally plausible or formally optimal, but whether it lets the user predict outcomes, avoid errors, and recover when things go wrong.

One way to formalize this is to distinguish the true system transition structure T from the user's internal approximation $\hat{T}_u$. Usability then depends not on full equality $\hat{T}_u = T$, but on low task-relevant divergence:

$$D_{\text{task}}(T, \hat{T}_u) \approx 0.$$

An interface succeeds when the user's conceptual model tracks the portions of system behavior needed for prediction and control. This applied tradition is still relevant because it reminds us that world models can be evaluated pragmatically. A model may be incomplete or even partly inaccurate, yet still useful for action in a given domain.

6.2.1 Consequences and implications

The main consequence here is better human performance in complex systems. Good user mental models made it possible to predict interface behavior, transfer knowledge across tasks, detect anomalies, and recover from errors without exhaustive trial-and-error. At a systems level, this perspective also clarified why apparently small representational mismatches between system and user can produce large downstream failures in safety, trust, and coordination.

Section Summary: Shared and Applied World Models

Main definitions. Shared and applied world models are representations used by groups, institutions, designers, or users to coordinate action, understand systems, and guide intervention in complex environments.

Main features.

- Stock-flow and feedback-loop structure in systems dynamics.
- Negotiated or shared causal representations across agents and stakeholders.
- User-facing conceptual approximations of technical systems in HCI.

Main achievements.

- Better policy design in the presence of delay, feedback, and emergence.
- Better coordination through explicit comparison of stakeholder models.
- Better usability, predictability, and recovery in human-system interaction.

7 Foundation World Models in Contemporary AI

Since 2024, a new usage of "world model" has become prominent in AI: the term increasingly denotes a foundation-scale simulator or predictive representation trained from massive video, spatial, game, and robotic data. The important shift is not only scale. It is a shift in what the model is expected to do. A contemporary foundation world model is evaluated not merely by one-step prediction loss or visual realism, but by whether it supports physical reasoning, action-conditioned rollout, agent training, synthetic data generation, sandboxing, robot planning, and robust interaction under interventions.

Formally, one can view this family as learning a broad conditional process

$$\hat{p}_\theta(x_{t+1:t+H}, z_{t+1:t+H}, r_{t:t+H-1} \mid c, x_{\le t}, a_{t:t+H-1}),$$

where x_t may include video frames, language, spatial tokens, proprioception, or other observations; z_t is a latent model state; a_t is an action, control input, or latent intervention; r_t is an optional task-value signal; and c is a conditioning context such as a prompt, image, goal, map, or task instruction. This template is deliberately broader than the model-based RL template of Section 4. A foundation world model may be trained without rewards, without explicit actions, and without a single downstream task. But it becomes a world model in the stronger intervention-oriented sense only when its latent or generative structure can be queried as an action-sensitive surrogate for the domain.

Figure 7 summarizes the contemporary landscape. The main branches are latent predictive models and JEPAs, generative interactive simulators, and embodied-agent training loops. Their common ambition is to turn large-scale sensory and interaction data into a model space that can support planning, sandboxing, robotics, and evaluation.

7.1 From Latent RL Models to Foundation World Models

The immediate technical precursor is the learned latent model of model-based reinforcement learning. Dreamer-style agents learn a recurrent state-space model with an encoder, a latent transition model, and prediction heads for rewards, values, and sometimes observations:

$$q_\phi(z_t \mid o_{\le t}, a_{<t}), \qquad p_\theta(z_{t+1} \mid z_t, a_t), \qquad p_\theta(r_t, o_t \mid z_t, a_t).$$

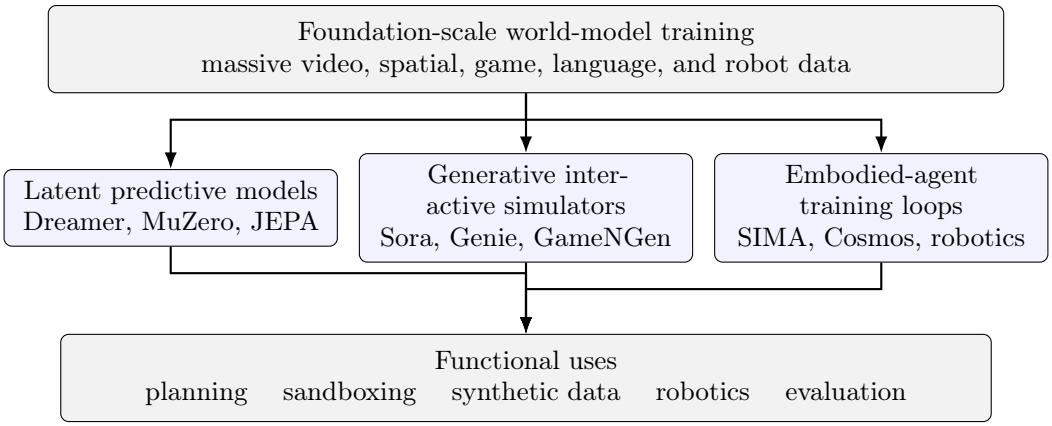

Figure 7: Three branches of contemporary foundation world models. The branches differ in training objective and interface, but converge on action-sensitive uses: planning, sandboxing, synthetic data generation, robotics, and functional evaluation.

The policy can then be optimized by imagined rollouts,

$$J(\pi) = \mathbb{E}_{\hat{p}_\theta, \pi}\left[\sum_{k=0}^{H-1} \gamma^k r_{t+k}\right],$$

rather than by interacting with the environment for every candidate trajectory [19, 20]. MuZero sharpens the point: the learned model need not reconstruct pixels if its latent dynamics preserve the information needed for policy and value prediction [21]. In both cases the model is task-specific, reward-shaped, and usually trained within a relatively narrow environment distribution.

Foundation world models generalize this pattern along three dimensions. First, the data distribution becomes broad: web video, game trajectories, robot demonstrations, simulations, and spatial captures replace one task or benchmark. Second, the model becomes multimodal: it may condition on text, images, actions, maps, goals, or proprioceptive streams. Third, the desired interface becomes more generative and controllable. The mathematical shift is from unconditional or weakly conditioned video generation,

$$\hat{p}_\theta(x_{1:T} \mid c),$$

to action-sensitive simulation,

$$\hat{p}_\theta(x_{t+1:t+H} \mid x_{\leq t}, a_{t:t+H-1}, c),$$

where actions may be explicit controls, joystick commands, robot actions, language instructions, or inferred latent actions.

OpenAI's Sora report made this shift visible by arguing that large video generation models may become "general purpose simulators of the physical world" [25]. Technically, the claim is plausible because video prediction forces a model to compress persistent objects, geometry, motion, occlusion, and many regularities of ordinary physics. But it is also incomplete. A video generator can produce plausible futures while failing at intervention semantics. The stronger world-model claim requires counterfactual controllability: if a_t changes while the context is held fixed, the resulting distribution over futures should change in ways that respect the causal structure of the domain. Thus the foundation-model version of a world model is not simply a bigger Dreamer or a prettier video predictor. It is a broad prior over dynamics that becomes useful when it can be connected to actions, goals, and downstream decision procedures.

7.2 The JEPA Line: LeCun, Meta, and Predictive Latent Representations

The Joint-Embedding Predictive Architecture (JEPA) line offers a second route to foundation world models. LeCun's proposal for autonomous machine intelligence places world modeling at the center of intelligent behavior, but argues that the relevant prediction should occur primarily in representation space rather than pixel space [26]. The key reason is selective abstraction. Pixels contain many details that are expensive to predict and often irrelevant for action. A useful world model should predict the latent variables that matter for object permanence, affordances, geometry, motion, and goal-directed control.

A generic JEPA objective can be written as

$$z_y = E_\theta(y), \qquad \hat{z}_y = P_\phi(E_\theta(x), m), \qquad \mathcal{L}_{\text{JEPA}} = d(\hat{z}_y, \text{sg}(z_y)),$$

where x is the visible context, y is a masked or future target, m describes the mask or prediction query, E_θ is an encoder, P_ϕ is a predictor, d is a representation-space discrepancy, and sg denotes a stop-gradient target. Unlike reconstruction-based masked autoencoding, the loss does not require the model to reproduce all low-level target pixels. It requires the model to infer a representation of the hidden or future content.

I-JEPA instantiates this idea for images by predicting representations of masked image regions from surrounding context [27]. V-JEPA extends the principle to video, where the target representation contains temporal structure and the model must infer what is likely to happen across space and time [28]. In world-model terms, the important commitment is that prediction in latent space can be more action-relevant than prediction in sensory space. The model is not rewarded for memorizing texture noise; it is rewarded for organizing features that make hidden and future scene structure predictable.

V-JEPA 2 makes the world-model interpretation more explicit. It combines large-scale self-supervised video training with a small amount of robot data and demonstrates planning for physical robot arms by evaluating candidate action sequences in the learned representation space [29]. A simplified planning rule is

$$a^*_{t:t+H-1} \in \arg\min_{a_{t:t+H-1}} d(F_\theta(z_t, a_{t:t+H-1}), z_g),$$

where z_t is the current visual representation, z_g is a goal representation, and F_θ rolls the representation forward under candidate actions. The model need not generate photorealistic futures to be useful. It only needs a latent geometry in which action-conditioned rollouts can be compared with goals.

V-JEPA 2.1 extends this direction toward dense spatial features, improving the suitability of representation-space prediction for tasks such as grasping, navigation, and spatially precise robot control [30]. This is conceptually important because many embodied tasks require both semantic abstraction and local metric detail. A purely global embedding may know that an object is present, while a dense feature field can also support where-to-move and where-to-contact decisions. The JEPA line therefore illustrates a distinctive thesis about foundation world models: understanding may be better measured by the structure of the predictive latent space than by the pixel-level realism of generated samples.

7.3 Generative Interactive Worlds: Genie, GameNGen, and Real-Time Simulation

A different branch treats the world model as a generative interactive environment. Here the model is valuable because it can synthesize an evolving world in response to user or agent actions:

$$x_{t+1} \sim G_\theta(x_{\le t}, a_t, c).$$

The challenge is not only to predict the next frame, but to maintain consistency, controllability, and real-time responsiveness over long horizons. For an interactive model, latency and coherence become part of the technical specification. If the target environment runs at step time Δt_{env}, the model must satisfy something like

$$\Delta t_{\text{model}} \leq \Delta t_{\text{env}}$$

while keeping accumulated rollout error small enough that the generated world remains navigable.

DeepMind's Genie line is central to this branch. Genie 1 is an 11-billion-parameter foundation world model trained from unlabeled internet videos and designed to generate action-controllable interactive environments [31]. A key technical step is latent action inference. When internet videos do not provide controls, the model can learn a discrete or continuous latent action code α_t from frame-to-frame changes,

$$q_\phi(\alpha_t \mid x_t, x_{t+1}), \qquad p_\theta(x_{t+1} \mid x_{\leq t}, \alpha_t, c).$$

This lets the model convert passive video into a controllable simulator-like representation. Genie 2 extends the ambition to 3D action-controllable world generation, where a prompt or image can specify a scene and user actions can drive the unfolding environment [32]. Genie 3 pushes the interaction constraint further, aiming at real-time world generation at 24 frames per second and 720p resolution over multi-minute episodes [33].

GameNGen provides a more focused demonstration of the same principle: a diffusion model can serve as a neural game engine when trained to generate the next game frame conditioned on recent frames and actions [34]. The significance is not that a particular game is solved by video prediction. It is that the rendering and transition function of an interactive world can be approximated by a learned generative model. This makes the boundary between simulator, environment, and model less sharp. In classical model-based RL, the simulator is the external ground truth and the world model approximates it. In neural game engines and generative worlds, the learned model itself becomes the environment in which agents or humans can act.

The main technical limitation is compounding distribution shift. Interactive rollouts visit states selected by the user or agent, not only states sampled from the training videos. If d_t denotes the rollout-state distribution induced by the learned model and policy, while d_t^* denotes the real environment distribution, long-horizon reliability depends on controlling a divergence such as

$$\sum_{t=1}^{H} D(d_t \,\|\, d_t^*)\,.$$

Visual fidelity can hide this problem. A generated world may look plausible while silently violating object permanence, action effects, conservation constraints, or task-relevant geometry.

7.4 World Models as Training Grounds for Agents

Foundation world models also matter because they can become training grounds for agents. The agent does not merely ask the model for predictions; it lives inside model-generated or model-augmented environments. This connects directly to the sandboxing interpretation of world models: a world model is useful when it provides an interface through which an agent can safely, cheaply, and repeatedly experience counterfactual consequences before acting in the real world [18].

The SIMA line makes this connection explicit. SIMA was introduced as a generalist agent trained to operate through a generic visual, language, and action interface across many 3D virtual environments [35]. Abstractly, an instruction-conditioned agent can be written as

$$a_t \sim \pi_\psi(a_t \mid o_{\leq t}, \ell, h_t),$$

where ℓ is a natural-language instruction and h_t is the agent's internal history state. The agent's objective is not tied to one game engine or one reward specification, but to robustly following instructions across environment distributions:

$$J(\pi_\psi) = \mathbb{E}_{e\sim\mathcal{E},\,\tau\sim M_e,\pi_\psi}\left[\sum_t R_e(o_t, a_t, \ell)\right].$$

Here $\mathcal{E}$ is a distribution over worlds and M_e is the transition process for one world. Generalization depends on whether the agent learns reusable perceptual, linguistic, and motor abstractions rather than memorizing environment-specific affordances.

SIMA 2 strengthens the loop by combining an agent based on Gemini with Genie 3-generated worlds for generalization and self-improvement [36]. The resulting pattern is cyclic:

$$\begin{aligned}\text{world model} &\to \text{agent experience} \to \text{policy improvement}\\ &\to \text{new tasks and rollouts} \to \text{world-model or curriculum refinement.}\end{aligned}$$

This is a new role for world models relative to earlier model-based RL. The model is not only a private latent predictor inside a single agent. It becomes a shared training substrate, curriculum generator, evaluation space, and source of hard cases. If the world model can generate controlled variations of scene, physics, layout, goals, and distractors, then it can test an agent's invariances more systematically than a fixed benchmark.

This also changes the evaluation question. The relevant metric is no longer only whether the model predicts a held-out video clip. It is whether agents trained or evaluated in the model acquire capabilities that transfer:

$$\Delta_{\text{transfer}} = J_{\text{real}}(\pi_{\text{trained in } M}) - J_{\text{real}}(\pi_{\text{baseline}}).$$

A foundation world model is useful as a training ground only if Δ_{transfer} is positive under meaningful real or independently simulated tasks. Otherwise it may produce attractive but self-contained worlds that overfit agents to the model's artifacts.

7.5 Physical AI, Robotics, and Spatial Intelligence

The robotics and "physical AI" branch emphasizes world models as spatially grounded simulators for embodied systems. NVIDIA Cosmos is the clearest recent example: a family of open world foundation models intended for robot and autonomous-vehicle simulation, synthetic data generation, scenario expansion, and downstream fine-tuning [37]. The motivation is practical. Robots and autonomous vehicles require enormous coverage over rare events, long-tail scenarios, contact dynamics, lighting variation, weather, occlusions, and human behavior. Collecting all such cases in the real world is costly, slow, and sometimes unsafe.

In this setting, a world foundation model acts as a data generator and counterfactual scenario engine. If $\mathcal{D}_{\text{real}}$ is the available real dataset, the model produces synthetic trajectories

$$\tilde{x}^{(i)}_{1:T} \sim \hat{p}_\theta(x_{1:T} \mid c_i, a^{(i)}_{1:T}), \qquad \mathcal{D}_{\text{train}} = \mathcal{D}_{\text{real}} \cup \{(c_i, a^{(i)}_{1:T}, \tilde{x}^{(i)}_{1:T})\}_{i=1}^N.$$

The goal is to improve downstream perception, planning, and control by increasing coverage of physically meaningful cases. The core risk is sim-to-real error:

$$\epsilon_{\text{sim2real}} = D(\hat{p}_\theta(x_{1:T} \mid c, a), p_{\text{real}}(x_{1:T} \mid c, a)),$$

especially in precisely those rare or safety-critical regimes where real data are sparse. Physical AI world models therefore require more than cinematic realism. They need calibrated geometry, dynamics, contact, agent behavior, and sensor statistics.

This branch also overlaps with commercial work on spatial intelligence. World Labs' Marble, for example, points toward interactive 3D world generation as a product direction rather than a conventional academic benchmark [38]. It should be cited more cautiously because public technical detail is thinner than in peer-reviewed papers. Still, it reflects the same trajectory: world models are increasingly framed as systems that can construct persistent, navigable, spatial environments rather than isolated images or clips.

7.6 Evaluation: Realism Is Not Yet Understanding

The central evaluation problem is that visual realism is not equivalent to world understanding. A model can synthesize convincing motion while failing physical counterfactuals, object permanence, mass and contact reasoning, or causal intervention tests. If the model is used only as a generator, perceptual quality metrics may be acceptable. If it is used as a world model, the evaluation must ask whether it supports the right queries.

One useful abstraction is to define a family of functional tests $\mathcal{Q}$ and score the model by query accuracy rather than sample appeal:

$$\mathrm{Eval}(M) = \mathbb{E}_{q\sim\mathcal{Q}}\left[S(A_M(q), A^*(q))\right],$$

where q may be a prediction, intervention, counterfactual, planning, or embodied-control query; $A_M(q)$ is the answer induced by the model; $A^*(q)$ is the target answer; and S is a task-specific score. A foundation world model should therefore be evaluated along several axes:

$$S_{\text{world}} = \lambda_{\text{f}} S_{\text{fidelity}} + \lambda_{\text{dyn}} S_{\text{dynamics}} + \lambda_{\text{act}} S_{\text{action}} + \lambda_{\text{cf}} S_{\text{counterfactual}} + \lambda_{\text{emb}} S_{\text{embodiment}}.$$

The weights depend on use case. For entertainment, fidelity and responsiveness may dominate. For robotics, action effects, geometry, contact, and transfer matter more.

Recent benchmarks and critiques move in this functional direction. Physics-IQ tests whether contemporary video and multimodal models exhibit physical understanding beyond visual plausibility, and reports substantial limitations [39]. "How Far is Video Generation from World Model?" argues that scaling video generation alone may not yield robust abstraction of physical laws [40]. EWMBench evaluates embodied world models in terms of action-conditioned prediction and task-relevant physical consistency [41]. WorldArena continues the shift toward interactive and functional evaluation, where the model is assessed by how well it supports agents and world-level tasks rather than by isolated frame quality [42].

The lesson is not that foundation world models are unimportant. It is that the name "world model" should be earned by intervention-sensitive competence. The strongest systems will combine broad generative priors, predictive latent representations, action-controllable rollout, spatial grounding, and evaluation protocols that test physical and causal structure. In that sense, the contemporary foundation-model literature does not replace older planning, control, robotics, and cognitive-science traditions. It scales them into a new regime where the open question is whether internet-scale predictive learning can be made reliable enough for real action.

Section Summary: Foundation World Models in Contemporary AI

Main definitions. A foundation world model is a large-scale predictive or generative surrogate trained on broad sensory, spatial, game, or robotic data and used to answer action-sensitive queries about future, hidden, or counterfactual structure.

Main features.

- Scale: broad training data rather than one task-specific environment.
- Interface: multimodal conditioning by text, images, goals, actions, spatial states, or robot controls.
- Objective: representation-space prediction, generative simulation, or agent-training utility.
- Evaluation: functional competence under planning, interaction, physical reasoning, and transfer, not visual realism alone.

Main achievements.

- Latent representation learning that supports planning without pixel reconstruction.
- Action-controllable generative environments and neural game-engine demonstrations.
- Agent-training loops that use generated worlds for curriculum, evaluation, and self-improvement.
- Synthetic-data and scenario-generation pipelines for robotics and autonomous systems.

8 Discussion

The survey supports a functional interpretation of "world model" rather than a format-specific one. Once the literature is organized into engineered, biological/cognitive, shared/applied, and contemporary foundation-model branches, the apparent heterogeneity becomes more tractable. Table 1 summarizes the main axes along which these traditions differ.

Axis	One pole	Other pole
Representation	Explicit, symbolic, interpretable action models	Learned, latent, distributed predictive models
Location	Internal cognitive or neural representation	External simulator or digital twin
Scope	World-centric, broad background structure	Task-centric, utility-preserving structure
Primary orientation	Observation and explanation	Intervention and control
Ownership	Individual model in a single agent	Shared or institutional model across teams or tools
Temporal emphasis	Static relational organization or event structure	Dynamics, transition, rollout, and feedback
Success criterion	Psychological plausibility or explanatory adequacy	Planning utility, control performance, or predictive accuracy

Table 1: Major axes that differentiate world-model traditions.

8.1 A functional family, not a single formalism

There is no single canonical world-model state space. In symbolic AI the modeled world is often a discrete action state graph; in control it is a dynamical system; in neuroscience it may be a latent relational or predictive space; in psychology it may be a situation model or possibility structure; in HCI it may be the user's working understanding of a device; and in contemporary generative AI it may be a multimodal latent process defined over video, text, action, and spatial tokens. The word "world" therefore ranges from literal physical environments to highly task-relative latent domains. What unifies the cases is not ontology, but role: the model serves as a manipulable surrogate that stands in for direct interaction with the domain.

This is why prediction alone is too weak as a general criterion. Many representations predict something, but not all of them organize a domain in a way that supports off-line reasoning. Cognitive maps support relational generalization; mental models support deduction and counterfactual thought; scripts support event comprehension; digital twins support monitoring and diagnosis; systems-dynamics models support policy discussion and organizational learning. Across the report, the best unifying characterization was not "a predictor of observations" but "a structured surrogate that preserves enough latent and dynamical organization for useful reasoning away from the world itself."

8.2 Action-sensitive surrogate reasoning is the strongest core

The strongest cross-disciplinary commonality is counterfactual competence under possible action. Classical planning makes this explicit by searching over action-conditioned transitions. Control and robotics optimize interventions under learned or engineered dynamics. Internal-model theories in neuroscience explain action by forward prediction and correction. Predictive-processing accounts reintroduce action through active inference. Mental-model theory studies how reasoners explore possibilities rather than simply extrapolating sensory streams. Even the most recent AI systems become interesting as world models not when they merely continue a sequence plausibly, but when they support intervention-sensitive queries about what would happen if the agent, user, or system did something else.

This point also sharpens the boundary of the term. A memory system, embedding space, classifier, or next-token predictor is not automatically a world model merely because it contains useful information about the environment. To count in the stronger sense used here, the representation must support queries about hidden state, possible futures, or intervention-dependent consequences. That criterion explains both why the concept applies across so many fields and why it should not be stretched to cover every predictive architecture.

8.3 Foundation AI is a new regime, not a new concept

The contemporary foundation-model literature changes the scale, modality, and interface of world modeling, but it does not erase the older conceptual structure. Large video, game, spatial, and robotics models extend the world-model idea toward internet-scale priors, multimodal conditioning, interactive simulation, synthetic-data generation, and agent-training environments. In that sense, recent systems are genuinely new. They expand what kinds of data can be absorbed into a surrogate model and what kinds of tasks can be supported by a single learned system.

At the same time, they inherit central ideas from older traditions: latent abstraction from model-based RL, intervention-sensitive simulation from planning and control, internal predictive structure from neuroscience, and externalized shared models from engineering and policy practice. The most important lesson of the contemporary literature is therefore not that visual realism has solved world modeling. It is that the longstanding question has reappeared at a new scale. The

relevant evaluation target is no longer only one-step prediction loss or perceptual fidelity, but whether the model supports planning, embodied transfer, physical reasoning, and robust interaction under intervention. Realism is useful; it is not yet understanding.

8.4 Implications for future use of the term

One practical implication of the survey is that claims about world models should be made more explicitly. Any serious use of the term should specify at least the modeled state or domain, the update mechanism, the intervention semantics, the class of supported queries, and the success criterion. Once those quantities are stated, the literature becomes easier to organize and compare. Without them, the phrase risks collapsing into a vague label for almost any model with broad predictive aspirations.

The broader implication is constructive rather than merely terminological. "World model" marks a recurring strategy for intelligence: shift part of the burden of inference, counterfactual evaluation, and coordination from the world itself into a surrogate space that can be updated, queried, and manipulated. Different disciplines instantiate that strategy differently, but the architectural role is stable. Seen in that light, symbolic action theories, control-theoretic dynamics models, latent imagination models, cognitive maps, internal models, mental models, digital twins, and foundation-scale generative systems are not isolated curiosities. They are neighboring realizations of one enduring idea about how intelligent systems escape purely reactive behavior.

References

[1] K. J. W. Craik. *The Nature of Explanation*. Cambridge University Press, 1943. Open Library: https://openlibrary.org/books/OL26537526M/The_nature_of_explanation.

[2] Plato. *Republic*, Book VII (514a–520a), especially the allegory of the cave. Standard online text: https://web.stanford.edu/class/ihum40/cave.pdf.

[3] Zhuangzi. *Zhuangzi*, Inner Chapters, Chapter 2 ("Discussion on Making All Things Equal"), especially the butterfly dream passage. Online text: https://en.wikisource.org/wiki/Chuang_Tz%C5%AD_(Giles)/Chapter_2.

[4] Aristotle. *On the Soul* (*De Anima*), Book III, especially the discussion of *phantasia*. Internet Classics Archive: https://classics.mit.edu/Aristotle/soul.3.iii.html.

[5] Aristotle. *On Dreams*. Internet Classics Archive: https://classics.mit.edu/Aristotle/dreams.html.

[6] Sextus Empiricus. *Outlines of Pyrrhonism*. Online Philosophy edition: https://onlinephilosophy.org/books/outlines-pyrrhonism.

[7] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4*, 1969. HTML version: https://www-formal.stanford.edu/jmc/mcchay69/mcchay69.html.

[8] R. E. Fikes and N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3–4):189–208, 1971. https://www.sciencedirect.com/science/article/pii/0004370271900105.

[9] P. J. Hayes. *The Naive Physics Manifesto*. In D. Michie, editor, *Expert Systems in the Micro-Electronic Age*. Edinburgh University Press, 1979. Google Books: https://books.google.com/books/about/The_Naive_Physics_Manifesto.html?id=Hco9HAAACAAJ.

[10] R. S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991. https://doi.org/10.1145/122344.122377.

[11] E. Davis and G. Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015. https://cacm.acm.org/research/commonsense-reasoning-and-commonsense-knowledge-in-artificial-intelligence/.

[12] S. Levine, C. Finn, T. Darrell, and P. Abbeel. Learning complex neural network policies with trajectory optimization. In *Proceedings of ICML*, 2014. https://proceedings.mlr.press/v32/levine14.html.

[13] C. Finn and S. Levine. Deep visual foresight for planning robot motion. In *Proceedings of ICRA*, 2017. https://arxiv.org/abs/1610.00696.

[14] F. Ebert, C. Finn, A. Dasari, A. X. Lee, and S. Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv*, 2018. https://arxiv.org/abs/1812.00568.

[15] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. DayDreamer: World models for physical robot learning. In *Proceedings of CoRL*, 2022. https://arxiv.org/abs/2206.14176.

[16] M. Grieves and J. Vickers. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In *Transdisciplinary Perspectives on Complex Systems*, 2017. https://event.asme.org/Events/media/library/resources/digital-twin/Digital-Twin-Transdisciplanary-Systems.pdf.

[17] T. Taniguchi, A. C. Varela, Y. Nagai, H. I. Christensen, A. K. Tanwani, M. Otte, and others. World models and predictive coding for cognitive and developmental robotics: Frontiers and challenges. *Advanced Robotics*, 37(13):780–806, 2023. https://doi.org/10.1080/01691864.2023.2225232.

[18] F. E. Rosas, A. Boyd, and M. Baltieri. AI in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability. *arXiv*, 2025. https://arxiv.org/abs/2504.04608.

[19] D. Ha and J. Schmidhuber. World models. *arXiv*, 2018. https://arxiv.org/abs/1803.10122.

[20] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. In *Proceedings of ICLR*, 2020. https://arxiv.org/abs/1912.01603.

[21] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020. https://www.nature.com/articles/s41586-020-03051-4.

[22] J. Lin, Y. Du, O. Watkins, D. Hafner, P. Abbeel, D. Klein, and A. Dragan. Learning to model the world with language. In *Proceedings of ICML*, 2024. https://proceedings.mlr.press/v235/lin24g.html.

[23] R. Ma, J. Zhang, M. Abdelrahman, K. Wang, X. Song, and others. WorldCoder, a model-based LLM agent: Building world models by writing code and interacting with the environment. In *Proceedings of NeurIPS*, 2024. https://arxiv.org/abs/2402.12275.

[24] W. Xia, D. Lu, D. Yan, Y. Zhang, and X. Huang. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. In *Proceedings of NeurIPS*, 2023. https://arxiv.org/abs/2305.14909.

[25] OpenAI. Video generation models as world simulators. Technical report, 2024. https://openai.com/research/video-generation-models-as-world-simulators.

[26] Y. LeCun. A path towards autonomous machine intelligence. OpenReview technical report, 2022. https://openreview.net/forum?id=BZ5a1r-kVsf.

[27] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of ICCV*, 2023. https://arxiv.org/abs/2301.08243.

[28] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv*, 2024. https://arxiv.org/abs/2404.08471.

[29] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, and others. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv*, 2025. https://arxiv.org/abs/2506.09985.

[30] L. Mur-Labadia, M. Muckley, A. Bar, M. Assran, K. Sinha, M. Rabbat, Y. LeCun, N. Ballas, and A. Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv*, 2026. https://arxiv.org/abs/2603.14482.

[31] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, and others. Genie: Generative interactive environments. *arXiv*, 2024. https://arxiv.org/abs/2402.15391.

[32] Google DeepMind. Genie 2: A large-scale foundation world model. Research blog, 2024. https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/.

[33] Google DeepMind. Genie 3: A new frontier for world models. Research blog, 2025. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/.

[34] D. Valevski, Y. Leviathan, M. Arar, and S. Fruchter. Diffusion models are real-time game engines. *arXiv*, 2024. https://arxiv.org/abs/2408.14837.

[35] Google DeepMind. A generalist AI agent for 3D virtual environments. Research blog and technical report, 2024. https://deepmind.google/discover/blog/sima-generalist-ai-agent-for-3d-virtual-environments/.

[36] Google DeepMind. SIMA 2: An agent that plays, reasons, and learns with you in virtual 3D worlds. Research blog and technical report, 2025. https://deepmind.google/en/blog/sima-2-an-agent-that-plays-reasons-and-learns-with-you-in-virtual-3d-worlds/.

[37] NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv*, 2025. https://arxiv.org/abs/2501.03575.

[38] World Labs. Marble and spatial intelligence. Product and research website, 2025. https://www.worldlabs.ai/.

[39] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos. Do generative video models understand physical principles? *arXiv*, 2025. https://arxiv.org/abs/2501.09038.

[40] B. Kang, Y. Yue, R. Lu, Z. Lin, Y. Zhao, K. Wang, G. Huang, and J. Feng. How far is video generation from world model: A physical law perspective. In *Proceedings of ICML*, 2025. https://arxiv.org/abs/2411.02385.

[41] H. Yue, S. Huang, Y. Liao, S. Chen, P. Zhou, L. Chen, M. Yao, and G. Ren. EWMBench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv*, 2025. https://arxiv.org/abs/2505.09694.

[42] Y. Shang, Z. Li, Y. Ma, W. Su, X. Jin, Z. Wang, L. Jin, X. Zhang, Y. Tang, H. Su, and others. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv*, 2026. https://arxiv.org/abs/2602.08971.

[43] E. C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. https://doi.org/10.1037/h0061626.

[44] J. O'Keefe and J. Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, 1971. https://doi.org/10.1016/0006-8993(71)90358-1.

[45] R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, and Y. Niv. Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2):267–279, 2014. https://doi.org/10.1016/j.neuron.2013.11.005.

[46] N. W. Schuck, M. B. Cai, R. C. Wilson, and Y. Niv. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91(6):1402–1412, 2016. https://doi.org/10.1016/j.neuron.2016.08.019.

[47] K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20:1643–1653, 2017. https://doi.org/10.1038/nn.4650.

[48] T. E. J. Behrens, T. Muller, J. C. R. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018. https://doi.org/10.1016/j.neuron.2018.10.002.

[49] J. C. R. Whittington, T. Muller, S. Mark, G. Chen, C. Barry, N. Burgess, and T. E. J. Behrens. The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020. https://doi.org/10.1016/j.cell.2020.10.024.

[50] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995. https://doi.org/10.1126/science.7569931.

[51] R. C. Miall and D. M. Wolpert. Forward models for physiological motor control. *Neural Networks*, 9(8):1265–1279, 1996. https://doi.org/10.1016/S0893-6080(96)00035-4.

[52] D. M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7–8):1317–1329, 1998. https://doi.org/10.1016/S0893-6080(98)00066-5.

[53] M. Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9(6):718–727, 1999. https://doi.org/10.1016/S0959-4388(99)00028-8.

[54] R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999. https://doi.org/10.1038/4580.

[55] K. Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456):815–836, 2005. https://doi.org/10.1098/rstb.2005.1622.

[56] K. Friston and S. Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B*, 364(1521):1211–1221, 2009. https://doi.org/10.1098/rstb.2008.0300.

[57] K. Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010. https://doi.org/10.1038/nrn2787.

[58] A. Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. https://doi.org/10.1017/S0140525X12000477.

[59] P. N. Johnson-Laird. Mental models in cognitive science. *Cognitive Science*, 4(1):71–115, 1980. https://doi.org/10.1016/S0364-0213(81)80005-5.

[60] P. N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010. https://doi.org/10.1073/pnas.1012933107.

[61] S. Khemlani, A. K. Barbey, and P. N. Johnson-Laird. Causal reasoning with mental models. *Frontiers in Human Neuroscience*, 8:849, 2014. https://doi.org/10.3389/fnhum.2014.00849.

[62] R. C. Schank and R. P. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates, 1977. Google Books: https://books.google.com/books/about/Scripts_Plans_Goals_and_Understanding.html?id=91rp8Jf2q74C.

[63] R. P. Abelson. Psychological status of the script concept. *American Psychologist*, 36(7):715–729, 1981. https://doi.org/10.1037/0003-066X.36.7.715.

[64] R. A. Zwaan, M. C. Langston, and A. C. Graesser. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5):292–297, 1995. https://doi.org/10.1111/j.1467-9280.1995.tb00513.x.

[65] R. A. Zwaan and G. A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185, 1998. https://doi.org/10.1037/0033-2909.123.2.162.

[66] L. W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660, 1999. https://doi.org/10.1017/S0140525X99002149.

[67] C. M. Atance and D. K. O'Neill. Episodic future thinking. *Trends in Cognitive Sciences*, 5(12):533–539, 2001. https://doi.org/10.1016/S1364-6613(00)01804-0.

[68] D. L. Schacter, D. R. Addis, and R. L. Buckner. Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, 8:657–661, 2007. https://doi.org/10.1038/nrn2213.

[69] D. A. Norman. Some observations on mental models. In D. Gentner and A. L. Stevens, editors, *Mental Models*. Lawrence Erlbaum Associates, 1983. Book page: https://www.routledge.com/Mental-Models/Gentner-Stevens/p/book/9781315802725.

[70] W. B. Rouse and N. M. Morris. On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, 100(3):349–363, 1986. https://doi.org/10.1037/0033-2909.100.3.349.

[71] P. M. Senge. *The Fifth Discipline: The Art and Practice of the Learning Organization*. Doubleday/Currency, 1990. Publisher page: https://www.penguinrandomhouse.com/books/163984/the-fifth-discipline-by-peter-m-senge/.

[72] J. K. Doyle and D. N. Ford. Mental model concepts for system dynamics research. *System Dynamics Review*, 14(1):3–29, 1998. Citation context: https://davidnford.engr.tamu.edu/wp-content/uploads/sites/83/2017/03/5.-System-Dynamics-Methodology.pdf.

Exercises on World Models

Companion Handout for "World Models Across Disciplines"

April 23, 2026

Thirty-Minute Exercise Set

Exercise 1: Boundary Test

Prompt. Consider the following systems:

- a weather forecast model,
- a recommender system,
- a robot simulator used for planning,
- a large language model answering questions about physics,
- a digital twin of a factory.

Task.

1. Decide which of these count as world models and which do not.
2. Give one criterion of your own for making that judgment.
3. Identify one case that is genuinely borderline and explain why.

Aim. Test the boundaries of the concept.

Exercise 2: Reverse the Central Claim

Prompt. The report argues that world models should be understood primarily by their function rather than by their representational format.

Task.

1. Argue the opposite position as strongly as you can.
2. Explain why representational format might matter after all.
3. Give one example of a system that would count as a world model under a purely functional definition but should, in your view, be excluded.

Aim. Develop charitable disagreement.

Exercise 3: Design a Failure Case

Prompt. Invent a system that looks impressive and may even be called a world model, but fails in an important way.

Task.

1. Describe the system in two or three sentences.
2. Explain what it does well.
3. Explain what it cannot do, choosing one of the following failure modes:
   - it predicts well but cannot support intervention,
   - it generates realistic outputs but has no causal grip,
   - it works in training settings but collapses under distribution shift,
   - it supports explanation but not control.
4. State what this failure reveals about how world models should be evaluated.

Aim. Shifts attention from outputs to criteria such as intervention, robustness, and action-guidance.

Exercise 4: Cross-Disciplinary Translation

Prompt. Take one idea associated with world models in one field and translate it into another field.

Possible starting points.

- What would a "digital twin" mean in psychology?
- What would a "mental model" mean in robotics?
- What would "predictive coding" look like in management science?
- What would a "cognitive map" amount to in human-computer interaction?

Task.

1. State what is gained by the translation.
2. State what is distorted or lost.
3. Decide whether the translation is illuminating or misleading overall.

Aim. See both the power and the danger of cross-disciplinary analogies.

Exercise 5: Realism Versus Action

Prompt. Which matters more for calling something a world model: realism or action-guidance?

Task.

1. Take one side.
2. Defend it with one concrete example.
3. Address a counterexample in which the other side seems stronger.

Optional debate format. One student argues that a highly realistic simulator counts as a world model even if it is weak for control. Another argues that a crude but action-useful surrogate is the stronger case.

Aim. Articulate what the report treats as the strongest core of the concept.

Exercise 6: Concept Stress Test

Prompt. Imagine an AI company announces a new system and describes it as a "foundation world model."

Task.

1. Describe, in a sentence or two, what the system claims to do.
2. Explain why the label "world model" might be useful.
3. Explain why the label might also be misleading or inflated.
4. List two pieces of evidence you would require before accepting the claim.

Aim. Connect the report's conceptual framework to scientific standards, hype, and evaluation practice.

Optional Capstone Exercise

Title. Build a Definition That Excludes Something You Care About

Task.

1. Write a definition of "world model" in four or five sentences.
2. Your definition must include at least two examples from different disciplines.
3. It must exclude at least one tempting borderline case.
4. It must state one evaluation criterion and explain why that criterion matters.
5. Exchange definitions with another student and try to break each other's boundary using a counterexample.

Aim. Do conceptual design rather than passive summary.