World models across disciplines

Concepts, architectures, and foundation models

Fernando E. Rosas

April 24, 2026

Why this topic?

- "World model" is now central in AI — especially in model-based RL, robotics, multimodal agents, and foundation-scale simulators.
- But the phrase does not pick out one single object across disciplines.
- This reflects repeated convergence on a common architectural role.
- What is shared and what varies across disciplines? What should count as a good world model?

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.

Narrow modern-AI usage

- action-conditioned predictive model,
- usually learned from data,
- typically evaluated by planning or policy utility.

Broader cross-disciplinary usage

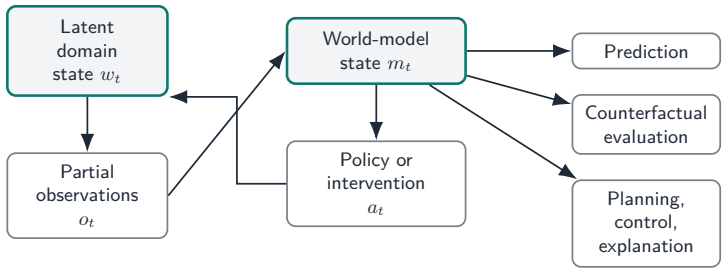
- cognitive maps,
- internal models,
- generative perceptual models,
- mental models,
- digital twins and shared causal models.

Minimal formal architecture

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

$$\hat{p}_\theta \left(\begin{matrix} w_{t+1:t+H}, \; o_{t+1:t+H}, \; r_{t:t+H-1} \\ \mid m_t, \; a_{t:t+H-1} \end{matrix} \right)$$

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



- w_t: latent domain state
- o_t: partial observations
- a_t: actions or interventions
- m_t: model state

The model state need not mirror the world one-for-one. It only needs to preserve the structure needed for the queries that matter.

Six axes of variation

World models vary along six recurring axes

Given vs learned	explicit hand-authored structure versus compact predictive structure learned from data
Internal vs external	representations inside agents or organisms versus external artifacts such as simulators and digital twins
Static vs dynamical	organizing states and event relations versus emphasizing rollout, feedback, and transition dynamics
Observational vs interventional	explaining sensory input versus asking what happens under chosen actions or perturbations
World-centric vs task-centric	broad background structure versus only what must be preserved for planning, control, or value estimation
Individual vs shared	private models for one agent versus negotiated models shared across groups, teams, or tools

What is not a world model?

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

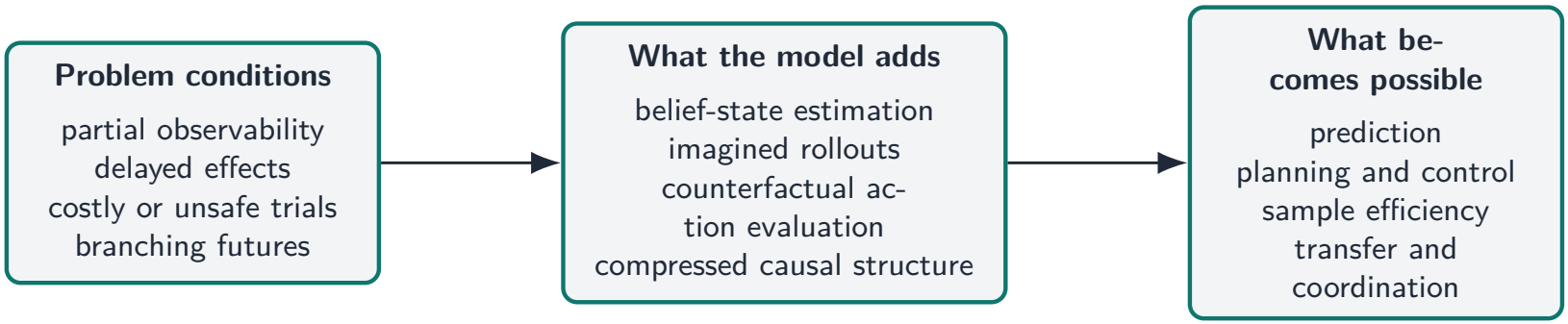
Three diagnostics

1. **Latent organization:** does it encode more than superficial associations?
2. **Intervention-sensitive querying:** can it be queried under hypothetical actions, goals, or perturbations?
3. **Multi-step surrogate reasoning:** can it support projection, comparison, explanation, or evaluation over trajectories?

Boundary condition

A memory buffer, lookup table, embedding, retriever, video generator, or next-token predictor may be useful without yet being a world model in the strong sense.

Why world models matter

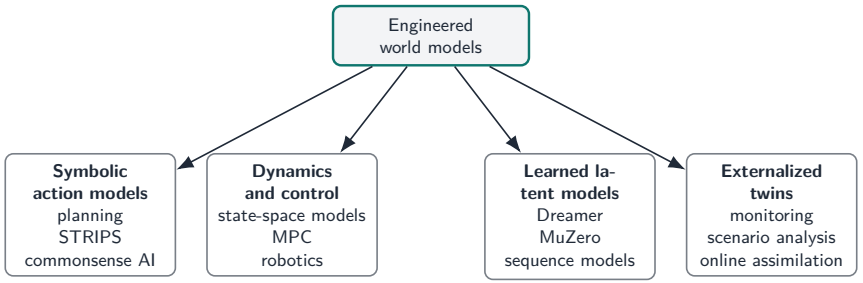


World models matter because they move part of reasoning out of expensive real interaction and into a manipulable surrogate space.

Engineered world models: the big picture

Definition

Engineered world models are artifact-level representations built for planning, control, search, monitoring, and policy optimization.



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

- symbolic action models,
- continuous dynamics models,
- learned latent simulators,
- externalized digital twins.

Symbolic action models and commonsense worlds

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

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

- Situation calculus and STRIPS treat the world as an explicit transition system.
- Commonsense AI extends the ontology with persistence, qualitative relations, and background physical structure.
- The frame problem becomes central: what changes, and what stays the same?

What this unlocked

- explicit planning,
- verifiable counterfactuals,
- reasoning about preconditions and side effects,
- structured commonsense inference.

Control, robotics, and digital twins

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

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

- In control, a world model is usually a forward dynamics or state-space model.
- In robotics, learned latent dynamics merge with model predictive control and trajectory optimization.
- In digital twins, the model is externalized and corrected by live data:

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

Consequence

These models moved expensive, dangerous, or irreversible trial-and-error into model space.

Learned latent world models

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

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

- Model-based RL made imagination a training signal, not just a metaphor.
- Dreamer shows policy learning from latent rollouts.
- MuZero shows a useful world model need not reconstruct the whole raw world.
- Recent extensions include sequence models, video-predictive models, multimodal embodied agents, and language-grounded planning.

Modern shift

The central question becomes decision utility: does the latent model preserve enough structure for planning, value estimation, transfer, and control?

Engineering family: so what?

Main definition

An engineered world model is a manipulable surrogate of state and dynamics built to support intervention.

Main features

- explicit or learned transitions,
- control-oriented evaluation,
- action-conditioned rollout,
- internal or external deployment.

Main achievements

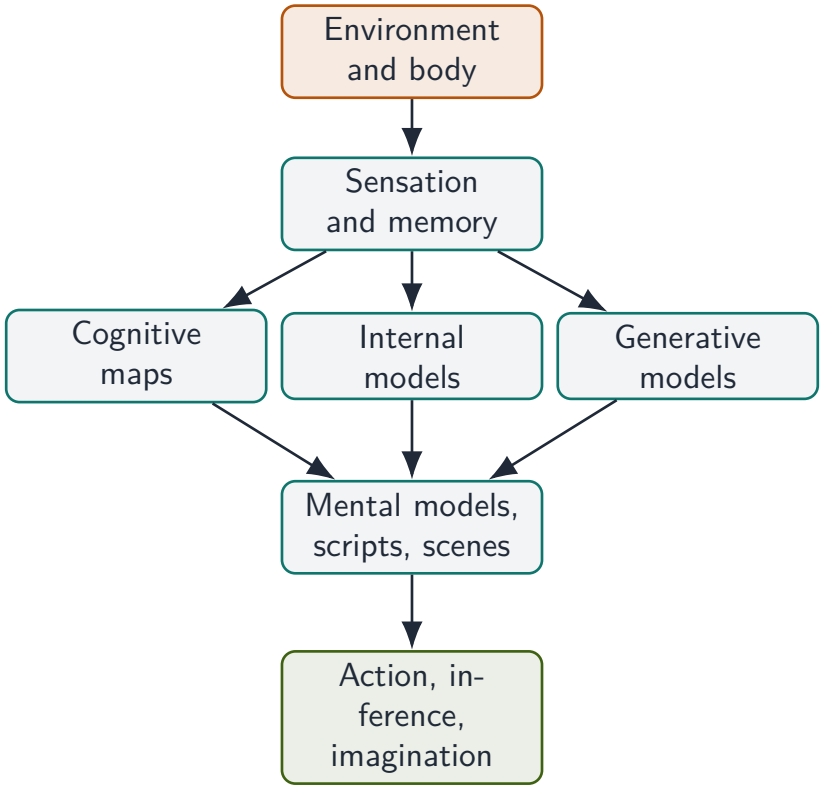
- automated plan synthesis,
- receding-horizon control,
- safer and more sample-efficient robot learning,
- counterfactual testing through twins and latent simulators.

Biological and cognitive world models

Shift in emphasis

Here the goal is not primarily engineered performance, but explaining how organisms perceive, predict, remember, reason, imagine, and act.

- There is no single biological analogue of the modern AI term.
- Several overlapping constructs do the work.
- The shared theme is latent structure that preserves future-relevant regularities.



Neuroscience I: maps, successor representation, replay

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

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

- Tolman, O'Keefe, and later hippocampal work treat the brain as representing latent environmental structure.
- The successor representation recasts a map as predictive future occupancy.
- Later work extends maps beyond literal space to task and relational structure.
- Replay suggests internally generated trajectories for planning and generalization.

Consequence: These models help explain navigation, replay, transfer, and generalization beyond simple stimulus-response learning.

Neuroscience II: internal models and predictive coding

Sensorimotor internal models

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

- Forward models predict sensory consequences of action.
- Inverse models support control and calibration.
- The payoff is stable control under delay and noise.

Predictive coding / active inference

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

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

- Perception becomes inference under a hierarchical generative model.
- Action can be treated as reducing expected future mismatch.

From prediction error to free energy

Derivation sketch

$$o = g(s) + \omega_o, \qquad s = f(\tilde{s}) + \omega_s$$

$$\omega_o \sim \mathcal{N}(0, \Sigma_o), \qquad \omega_s \sim \mathcal{N}(0, \Sigma_s)$$

$$\epsilon_o = o - g(s), \qquad \epsilon_s = s - f(\tilde{s})$$

$$-\log p(o, s) = \tfrac{1}{2}\epsilon_o^\top \Pi_o \epsilon_o + \tfrac{1}{2}\epsilon_s^\top \Pi_s \epsilon_s + C$$

$$\Pi_o = \Sigma_o^{-1}, \qquad \Pi_s = \Sigma_s^{-1}$$

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

If $q(s) \approx \delta(s - \mu)$, $\quad F[q] \approx \tfrac{1}{2} \sum_\ell \epsilon_\ell^\top \Pi_\ell \epsilon_\ell + C$

Toy example

$$o \sim \mathcal{N}(s, \sigma_o^2), \qquad s \sim \mathcal{N}(0, \sigma_s^2)$$

$$\epsilon_o = o - s, \qquad \epsilon_s = s$$

$$F(s) \approx \frac{(o - s)^2}{2\sigma_o^2} + \frac{s^2}{2\sigma_s^2} + C$$

$$\hat{s} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_o^2}\, o$$

- sensory data pulls $\hat{s}$ toward o
- the prior pulls $\hat{s}$ toward 0
- precision decides which term matters more

Prediction errors are the local mismatches; free energy is the global objective built from precision-weighted errors.

Psychology: mental models, scripts, situation models, prospection

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

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

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

- **Mental models:** structured possibility spaces for reasoning and counterfactuals.
- **Scripts / schemas / situation models:** event-structured expectations for comprehension.
- **Prospection:** constructive simulation of future scenes from memory traces and goals.

Biological / cognitive family: so what?

Main definition

These are internal representations posited to explain perception, action, memory, reasoning, and imagination.

Main features

- relational task spaces,
- sensorimotor prediction,
- hierarchical generative models,
- possibility- and scene-based simulation.

Main achievements

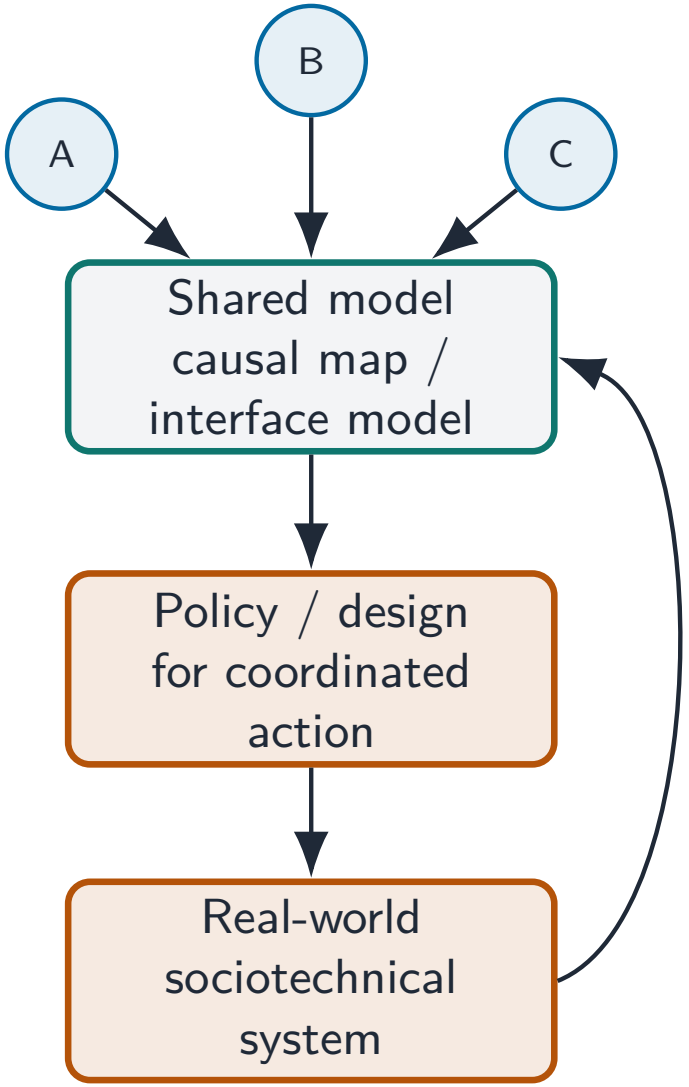
- explaining navigation and replay,
- explaining fast motor control under delay and noise,
- reframing perception as inference,
- explaining reasoning and prospection as structured simulation.

Shared and applied world models

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

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

- **Systems dynamics:** stocks, flows, delays, feedback loops.
- **Shared mental models:** negotiated causal pictures across stakeholders.



Shared / applied family: so what?

What these models made possible

- Better policy design under delay, feedback, and emergence.
- Better coordination through explicit comparison of stakeholder assumptions.
- Better human performance in complex interfaces and sociotechnical systems.
- A shift from individual cognition alone to group-level and institutional world modeling.

These traditions matter because they show that world models can be shared, externalized, negotiated, and only partially accurate while still being useful.

Four modern routes

Shared computational motif

learn model $\rightarrow$ imagine or predict futures $\rightarrow$ choose actions or generate trajectories

What now varies

- whether the model trains a policy,
- whether it supports search online,
- whether it first learns a predictive representation,
- whether it becomes an external simulator.

Most important axes from slide 4

- task-centric vs world-centric,
- observational vs interventional,
- internal vs external,
- individual vs shared.

Teaching claim

Ha and Dreamer, MuZero, JEPA, and Sora are in the same broad family, but they occupy different roles in the decision loop.

Modern AI world models

Family	What the model learns	How the model is used
Ha / Dreamer	latent dynamics, rewards, and policy for one task	imagined rollouts train a controller, actor, or critic; planning is largely compiled into the policy
MuZero	latent dynamics only insofar as reward, value, and policy structure are preserved	search queries the model online at decision time
JEPA / V-JEPA	predictive representations of what matters, rather than full sensory reconstruction	first a representation substrate; later attached to planning or robot control
Sora / Genie	large simulators of world evolution, sometimes with latent or explicit action interfaces	generate futures, synthetic worlds, or interactive environments; action use varies by system

The label "world model" alone is too coarse. The deeper question is which role the model plays in acting, searching, or simulating.

Ha and Dreamer: model trains the controller

$$z_t = E(o_t), \qquad h_{t+1} = f_\theta(h_t, z_t, a_t), \qquad J(\pi) = \mathbb{E}_{\hat{p}_\theta, \pi}\left[\sum_{t=0}^{H-1} \gamma^t \hat{r}_t\right]$$

- Ha and Schmidhuber learn a compressed latent state plus recurrent dynamics, then optimize a small controller on top.
- Dreamer turns the same motif into actor–critic RL: actor and critic learn mostly from imagined latent rollouts.
- The defining use of the model is **policy training**, not online tree search at every step.

Axis reading

Learned, **internal**, **task-centric**, **interventional**, **individual**, **dynamical**.

MuZero: model supports search

$$s_t^0 = h_\theta(o_{1:t}), \qquad (r_{k+1}, s_t^{k+1}) = g_\theta(s_t^k, a_{t+k}), \qquad (p_t^k, v_t^k) = f_\theta(s_t^k)$$

- MuZero does not try to reconstruct observations; it learns whatever latent structure is enough for reward, value, and policy targets.
- The model is queried by Monte Carlo tree search at decision time.
- So MuZero is close to Dreamer in architecture, but different in **how the model enters action selection**: search instead of a compiled actor.

Axis reading

learned, **internal**, **task-centric**, **interventional**, **individual**, and **dynamical**.

JEPA and V-JEPA: representation first

$$\hat{z}_y = P_\phi(E_\theta(x), m), \qquad \mathcal{L}_{\text{JEPA}} = d(\hat{z}_y, \text{sg}(z_y))$$
$$a^*_{t:t+H-1} \in \arg\min_{a_{t:t+H-1}} d(F_\theta(z_t, a_{t:t+H-1}), z_g)$$

- JEPA predicts embeddings rather than pixels.
- The bet is selective abstraction: preserve geometry, persistence, affordances, and other structure useful later.
- By itself this is often representation learning first; V-JEPA moves it toward planning and robot control.

Axis reading

learned, **internal**, more **world-centric** than MuZero, **interventional**, **individual**, and **dynamical**.

Sora and Genie: generative simulators

$$x_{t+1:t+H} \sim p_\theta(x_{t+1:t+H} \mid x_{\le t}, c), \qquad x_{t+1} \sim G_\theta(x_{\le t}, a_t, c)$$

- Sora makes the large-scale simulator ambition visible: broad world-centered video generation from prompts and context.
- Genie pushes closer to a world model in the strong sense by inferring actions and building controllable interactive worlds.
- The core risk is apparent realism without correct intervention semantics.

Axis reading

learned, **external**, **world-centric**, **interventional** for Genie, potentially **shared**, and strongly **dynamical**.

From internal model to external simulator

Agent-centric route

- Ha and Dreamer, MuZero, and most JEPA variants live inside one agent's decision loop.
- Success is better control, search, or value estimation.
- The dominant axes are internal, individual, and task-centric.

System-centric route

- Sora-, Genie-, Cosmos-, and digital-twin-like systems externalize the model into a simulator or shared tool.
- Success is broader: synthetic data, sandboxing, evaluation, coordination, and transfer.
- The axes shift toward external, shared, and more world-centric use.

Why this matters

The frontier is not just larger internal world models. It is also the externalization of model space into reusable environments for many agents, tasks, and users.

Axis map for the modern AI cases

Ha / Dreamer

Role: model trains the policy.
Axes: internal, task-centric, interventional, individual, learned, dynamical.

MuZero

Role: model supports search online.
Axes: internal, strongly task-centric, strongly interventional, individual, learned, dynamical.

JEPA

Role: model learns predictive abstraction first.
Axes: internal, more world-centric than MuZero, observational $\rightarrow$ interventional, individual, learned, dynamical.

Sora / Genie

Role: model becomes a simulator.
Axes: more external, more world-centric, Sora observational and Genie interventional, potentially shared, learned, dynamical.

Same motif, different axis positions: that is why these systems are related, but not interchangeable.

Five questions for any claimed world model

1. **What state is represented?** Reconstructed observation, predictive embedding, value-equivalent latent, or simulator state?
2. **How is that state updated?** Recurrence, search unroll, masked prediction, or generative rollout?
3. **How does the model enter action selection?** Learned actor, tree search, downstream planner, or external user?
4. **Which interventions actually matter?** Actions, prompts, latent controls, or physical perturbations?
5. **Where does it sit on the axes?** Internal or external, task-centric or world-centric, individual or shared?

Why this matters

The term becomes much less vague once the role of the model in the decision loop is made explicit.

Implications for modern AI

- Lumping Dreamer, MuZero, JEPA, and Sora together hides different commitments about state, action, and abstraction.
- Progress should not be measured by realism alone, nor by benchmark reward alone.
- Better evaluation asks whether the model preserves the distinctions needed for control, search, transfer, and counterfactual intervention.
- Older traditions still contribute design criteria: symbolic AI for explicit intervention structure, control for stability and constraints, neuroscience and psychology for predictive abstraction, and systems theory for shared external models.

Conclusion

One-sentence conclusion

"World model" names a family of surrogate spaces, not a single architecture.

- Ha and Dreamer train policies from imagination.
- MuZero supports search in a value-relevant latent space.
- JEPA learns predictive abstractions before full control is attached.
- Sora and Genie push toward external simulators and shared training worlds.
- The six axes explain why these systems are related, but not interchangeable.

Selected readings I

- Craik, *The Nature of Explanation* (1943)
- McCarthy and Hayes (1969); Fikes and Nilsson (1971)
- Hayes, "The Naive Physics Manifesto" (1979)
- Sutton, "Dyna" (1991)
- Ha and Schmidhuber, "World Models" (2018)
- Hafner et al., "Dreamer" (2020); Hafner et al., "DreamerV3" (2023)
- Schrittwieser et al., "MuZero" (2020)
- Grieves and Vickers on digital twins (2017)

Selected readings II

- Tolman (1948); O'Keefe and Dostrovsky (1971)
- Stachenfeld et al. on successor representations / predictive maps
- Behrens et al. on cognitive maps beyond space (2018)
- Wolpert, Miall, Kawato on internal models in motor control
- Rao and Ballard (1999); Friston (2005, 2009, 2010); Clark (2013)
- Johnson-Laird (1980, 2010); Khemlani et al. (2014)
- Schank and Abelson (1977); Zwaan and Radvansky (1998)
- Norman (1983); Rouse and Morris (1986); Senge (1990); Doyle and Ford (1998)

Selected readings III

- LeCun on autonomous machine intelligence and the JEPA line
- I-JEPA, V-JEPA, and V-JEPA 2 on predictive latent representations
- OpenAI on Sora as a possible general-purpose simulator
- Genie and GameNGen on controllable interactive worlds
- SIMA on agent training across many 3D environments
- NVIDIA Cosmos on physical AI world foundation models
- Physics-IQ, EWMBench, and WorldArena on evaluation beyond realism

Discussion