Toward Active World Modeling

ATOM Research Program v1.0.2

July 06, 2026

Author: Lancelot Da Costa

Context: The goal of the ATOM research community is to develop an **agent foundation model**: A machine learning model capable of (1) interaction (action, planning), (2) sample-efficient, continual and open-ended learning, (3) generalization and adaptation under distribution shifts, (4) alignment, safety and robustness.

Why: Current dominant approaches face critical limitations: neural net-based planning is largely amortized over training rather than adaptive at test time; LLMs are data-hungry and continual learning is a challenge; neural networks excel at interpolation but struggle to extrapolate under distribution shifts; and safety mechanisms remain brittle post-hoc patches.

How: We aim to jointly address these challenges through **active world modeling** under partial observability, where agents continually infer the states, parameters, and structural form of their world models while acting to maximize utility and information gain. Realizing this involves three interacting thrusts: (1) universal world modeling using hierarchical state-space models; (2) model discovery on hierarchical model spaces; and (3) planning for model refinement, safety and control. As we explain below, these foundations should advance the four criteria that define an agent foundation model.

When: As interest in world modeling accelerates alongside continued progress in LLMs, now is the opportune moment to develop this program that emphasizes active and partially observed settings, which are currently underexplored.

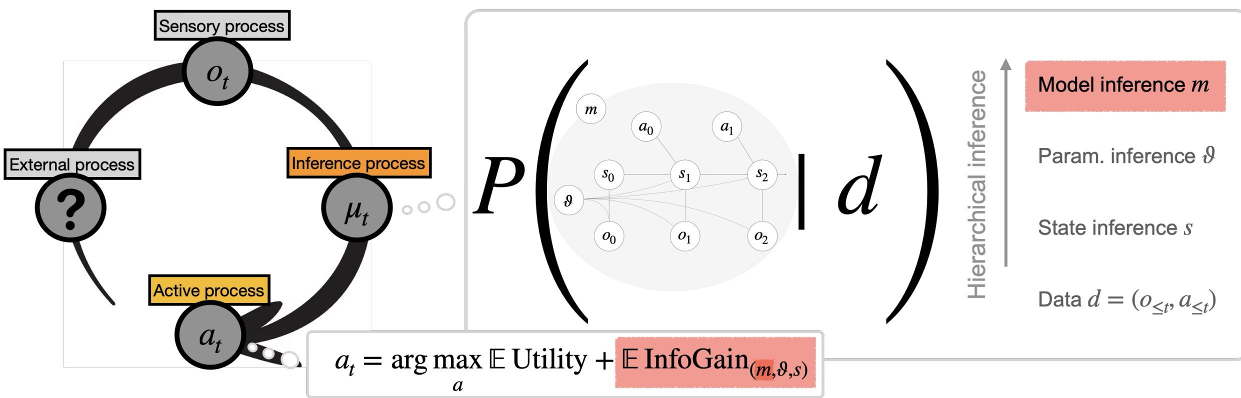

Figure 1: **Active World Modeling.** Agents perform hierarchical inference over heterogeneous spaces of state-space models (m), parameters (ϑ), and states (s) while planning to maximize expected information gain, and optionally a utility function—supporting few-shot scientific-like world modeling and goal achievement.

Proposed Approach: Active World Modeling

Our proposed approach develops the foundations of *active world modeling*: agents that jointly infer states, parameters, and structural form of their models while acting to gain information and maximize utility. Figure 1 summarizes this agenda: scalable hierarchical Bayesian inference over a heterogeneous space of structured world models, coupled with information-theoretic planning for informative data selection and control. Turning this conceptual picture into working agents raises three interlinked challenges for our research community: (1) *universal world modeling* for complex, partially observed environments; (2) *model discovery* on heterogeneous model spaces; and (3) *planning for model refinement and control* under uncertainty, distribution shifts and safety constraints. The remainder of this section outlines these three thrusts, which together are designed to support agents with enhanced sample-efficiency, calibrated uncertainty, adaptation under distribution shifts, and few-shot scientific-like world modeling and goal achievement.

Aim 1: Universal world modeling. We will explore hierarchies of discrete and continuous state-space models (e.g. extending POMDPs and rsLDS) as a unifying probabilistic backbone for complex, partially observed environments [1]. We will (1) develop efficient state and parameter variational inference methods for these hierarchies; (2) analyze the expressiveness of this model class mathematically through approximation theorems and empirically through multimodal generative modeling, identifying gaps; and (3) extend this model class as needed, e.g., to express latent non-Markovian SDEs [2]. Beyond algorithms, we will study identifiability, consistency, and statistical-computational tradeoffs. This thrust connects work in probabilistic modeling, variational inference, and probabilistic programming.

Aim 2: Model discovery on heterogeneous model spaces. World-modeling agents must reason not only about latent states and parameters, but also infer *which* models best explain their data. We will develop model discovery methods based on agent–environment interaction data through probabilistic inference over the model spaces of Aim 1. We will explore generative flow networks to encode posterior uncertainty [3], and priors grounded in core knowledge and minimum-description-length to support sample-efficient learning and generalization [4, 5]. We will develop information geometry and dynamical-systems theory over heterogeneous spaces of probabilistic models—inherently stratified spaces—yielding novel geometry-informed variational inference methods for model discovery with guarantees on approximation quality and robustness [6]. This thrust bridges nonparametric inference and high-dimensional model selection with causal representation learning and probabilistic program induction.

Aim 3: Planning for model refinement, safety and control. Given the state, parameter and model inferences from Aims 1–2, this thrust asks how agents should *act* to optimally refine their models while satisfying task, safety, and reliability constraints. We will develop information-theoretic planning schemes where agents select actions to maximize information gain about states, parameters, and models, subject to utility. We will extend work at the intersection of reinforcement learning and nonparametric Bayes [7] and previous work on information-theoretic curiosity-driven planning [8–10] with modern ideas in Bayesian experimental design [11]. Key foci include regret and sample complexity bounds under partial observability, robustness under distribution shifts, and characterizing regimes where sample complexity approaches human-level on few-shot modeling tasks (e.g. ARC-3). This thrust bridges Bayesian experimental design, active learning, and model-based planning.

Impact. These foundations support open-ended learning agents capable of autonomous scientific discovery, with enhanced sample efficiency, calibrated uncertainty that is essential for safety-critical applications, and adaptation under distribution shifts. They also support AI alignment research—framing alignment as modeling other agents' world models and utility functions, and incorporating those utilities as constraints in one's own [6].

Appendix: Key scientific background

Stochastic Processes and Bayesian Mechanics. The emerging field of *Bayesian mechanics*, a probabilistic mechanics that links multipartite stochastic processes with Bayesian inference and decision theory [12, 13], shows that states or paths on one side of a boundary can be described as optimizing approximate posterior distributions about states or paths on the other, consistently with variational inference. This provides a mathematical physics formalization of popular principles for describing biological intelligence (including the Bayesian brain hypothesis and free-energy principle), which describe the brain as making sense of the world through approximate Bayesian inference. This viewpoint grounds approximate Bayesian inference as a physics and cognitive science-inspired approach for building AI systems.

Active Bayesian inference. Building on Bayesian mechanics, the *active Bayesian inference* framework formalizes agents through variational Bayesian inference for perception and learning, and Bayesian decision theory and experimental design for planning [14]. Here, planning optimizes a combination of expected utility and information gain, incorporating uncertainty quantification and information-theoretic curiosity into reinforcement learning (RL). This framework is closely related to maximum-entropy RL, generalizing standard reward maximization objectives [15]. Analyses of existing robotics models illustrate how active Bayesian inference can enable safety-aware, robust, and adaptive behavior [16]. Independently, a group from MIT showed that agents following these principles reach human-level sample efficiency in low-dimensional domains [17].

Emerging foundations for world-modeling agents. To make these principles actionable in complex environments, recent work develops probabilistic foundations for *world-modeling agents* via structured world models: agents that represent environments as hierarchical, partially observed stochastic processes rather than monolithic autoregressive ones. Recent work (1) identified a heterogeneous, provably expressive, and yet searchable space of Bayesian world models, formalized as hierarchies of discrete and continuous state-space models that extend partially observed Markov decision processes (POMDPs) and recurrent switching linear Gaussian dynamical systems (rsLDS) [2, 1]; (2) developed efficient search algorithms on a model subclass, enabling multimodal generative modeling (video, sound) with Bayesian hierarchical models [18]; (3) proposed normative principles for world-modeling agents based on Bayesian inference on model spaces, information geometry, core knowledge and minimal description length priors, and Bayesian model reduction—which form the core of the approach proposed in this document [6]. This strand connects to broader efforts in probabilistic modeling, Bayesian inference and causal representation learning.

References

[1] L. Da Costa, S. Namjoshi, M. A. Ansari, and B. Schölkopf, "Natural building blocks for structured world models: Theory, evidence, and scaling," 2025. arXiv:2511.02091.

[2] L. Da Costa, "Toward universal and interpretable world models for open-ended learning agents," in *NeurIPS 2024 Workshop on Intrinsically Motivated and Open Ended Learning*, 2024.

[3] Y. Bengio, S. Lahlou, T. Deleu, E. J. Hu, M. Tiwari, and E. Bengio, "Gflownet foundations," *Journal of Machine Learning Research*, vol. 24, no. 210, pp. 1–55, 2023.

[4] L. Da Costa, N. Da Costa, and B. Schölkopf, "From overfitting to generalization: The need for good priors in model discovery." In preparation.

[5] T. Pouncy and S. J. Gershman, "Inductive biases in theory-based reinforcement learning," *Cognitive Psychology*, vol. 138, p. 101509, 2022.

[6] L. Da Costa, T. Gavenčiak, D. Hyland, M. Samiei, C. Dragos-Manta, C. Pattisapu, A. Razi, and K. Friston, "Possible Principles for Aligned Structure Learning Agents," *Neural Computation*, pp. 1–47, 2025.

[7] F. Doshi-velez, "The Infinite Partially Observable Markov Decision Process," in *Advances in Neural Information Processing Systems*, vol. 22, Curran Associates, Inc., 2009.

[8] K. Friston, L. Da Costa, D. Hafner, C. Hesp, and T. Parr, "Sophisticated inference," *Neural Computation*, vol. 33, no. 3, pp. 713–763, 2021.

[9] T. Champion, L. Da Costa, H. Bowman, and M. Grześ, "Branching time active inference: the theory and its generality," *Neural Networks*, 2022.

[10] A. Paul, N. Sajid, L. Da Costa, and A. Razi, "On efficient computation in active inference," *Expert Systems with Applications*, 2024.

[11] T. Rainforth, A. Foster, D. R. Ivanova, and F. B. Smith, "Modern Bayesian Experimental Design," 2023. arXiv:2302.14545.

[12] L. Da Costa, K. Friston, C. Heins, and G. A. Pavliotis, "Bayesian mechanics for stationary processes," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 477, no. 2256, p. 20210518, 2021.

[13] K. Friston, L. Da Costa, N. Sajid, C. Heins, K. Ueltzhöffer, G. A. Pavliotis, and T. Parr, "The free energy principle made simpler but not too simple," *Physics Reports*, vol. 1024, pp. 1–29, 2023.

[14] A. Barp, L. Da Costa, G. França, K. Friston, M. Girolami, M. I. Jordan, and G. A. Pavliotis, "Geometric methods for sampling, optimisation, inference and adaptive agents," in *Handbook of Statistics*, vol. 46, pp. 21–78, 2022.

[15] L. Da Costa, N. Sajid, T. Parr, K. Friston, and R. Smith, "Reward maximization through discrete active inference," *Neural Computation*, vol. 35, no. 5, pp. 807–852, 2023.

[16] L. Da Costa, P. Lanillos, N. Sajid, K. Friston, and S. Khan, "How active inference could help revolutionise robotics," *Entropy*, vol. 24, no. 3, p. 361, 2022.

[17] P. A. Tsividis, J. Loula, J. Burga, N. Foss, A. Campero, T. Pouncy, S. J. Gershman, and J. B. Tenenbaum, "Human-level reinforcement learning through theory-based modeling, exploration, and planning," *arXiv preprint*, 2021. arXiv:2107.12544.

[18] K. Friston, C. Heins, T. Verbelen, L. Da Costa, T. Salvatori, D. Markovic, A. Tschantz, M. Koudahl, C. Buckley, and T. Parr, "From pixels to planning: scale-free active inference," *Frontiers in Network Physiology*, 2025.