Position: Hierarchical World Models with Causal Curation for Generalizing Agents

Fei Dai*
Department of Psychology
University of California, Berkeley
Berkeley, CA 94720
f_d@berkeley.edu

Hanqi Zhou*
Cluster of Excellence Machine Learning
University of Tübingen
Tübingen, Germany
hanqi.zhou@uni-tuebingen.de

Alison Gopnik
Department of Psychology
University of California, Berkeley
Berkeley, CA 94720
gopnik@berkeley.edu

Abstract

Current reinforcement learning agents excel at solving narrow tasks yet struggle to generalize beyond the environments they were trained in, which is often attributed to a lack of causal understanding. While learning world models can provide foresight, they are often brittle, whereas pure causal discovery remains intractable in complex settings. In this position paper, we argue that the path to a general-purpose agent lies not in building a single, monolithic world model, but in actively *curating* a compact and transferable library of causal knowledge. We introduce a three-level hierarchical framework to formalize this idea. Low-level *Executors* and mid-level *Controllers* learn context-specific predictive world models for motor control and skill execution. At the highest level, a *Curator* uses counterfactual reasoning over imagined tasks to maintain causal models that are most likely to be useful for future generalization. This framework recasts agent intelligence, instead of "prediction", but as a proactive "curation" of causal knowledge, leading to more resource-efficient and robust generalization.

1 Introduction

A hallmark of intelligence is the ability to perform novel tasks with minimal supervision, a capability that in humans relies on rich mental representations of the world [33]. Yet a significant gap remains between the narrow expertise of current reinforcement learning (RL) agents and the adaptive, goal-directed behavior of biological agents [8]. Rather than restating that "world models help," this position paper offers a new perspective to fuse *intuition* and *reasoning* in lifelong learning: predictive world models learn fast and local imagination [18], structural causal models learn deliberated and global counterfactuals [36], both empowered by an LLM curator that proposes "where is valuable to start." Here, RL embodiment supplies interventional data; empowerment decides what to try next [26, 43]. The result is, like human learning, an accumulated and compact library of causal abstractions and skills that transfers across tasks [16, 34, 33].

*These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Bridging Language, Agent, and World Models for Reasoning and Planning.

This necessity arises from two constraints. First, any embodied agent is *resource-bounded* with finite computation and memory [27, 7]. Second, the *"big world hypothesis"* posits that for most problems of interest, the environment is orders of magnitude more complex than the agent itself [20, 9]. These constraints rule out two extremes: the agent cannot perform exhaustive, high-fidelity modeling (e.g., full Bayesian inference over all possibilities), and the designer cannot hand-craft a learning algorithm that explicitly anticipates every environment the agent may encounter. Instead, intelligence requires curating simplified, abstract, yet sufficiently relevant representations that capture the environment dynamics for effective planning and generalization.

To build such general-purpose agents, the field has gone through several paradigms. Causality has long been framed as the route to robust generalization in both RL [10, 36] and the cognitive sciences [12, 3, 48]: humans posit causal hypotheses, intervene in them, and reuse them across contexts. Developmental studies reveal that even young children poke and manipulate objects in ways that are maximally informative, allowing them to disentangle cause from effect [15, 13]. This human-centric perspective suggests a direction for artificial agents [32, 37, 24]—instead of solely relying on extrinsic rewards, an agent should be intrinsically motivated to gain information about the world's causal mechanisms. Yet in practice, without strong priors or carefully guided interventions, agents relying on observational or limited interaction data learn slowly and often fail to identify the correct causal structure [21, 4, 28]. Concurrently, the field pivoted to predictive world models, which support imagination-based planning and are formally necessary for multi-step goal generalization [34]. However, these models often overfit, making extrapolation brittle . While LLMs offer broad linguistic priors, their knowledge is fundamentally ungrounded and acausal, stemming from textual correlations rather than embodied interaction[40, 5, 46, 1, 29].

These limitations present a trilemma: a choice between the knowledge richness of large language models, the grounded learning of model-free RL, and the robust generalization promised by causal models. We propose a new perspective on building a general-purpose agent by integrating these paradigms into a cycle of hypothesis, experimentation, and validation. The agent actively curates a compact and robust library of causal mechanisms, unfolding in three stages. First, the agent leverages an LLM's broad prior knowledge to generate plausible tasks to test **causal hypotheses**, transforming open-ended discovery into a more constrained model selection problem [30, 26]. Second, acting as an **active experimentalist**, the agent designs and executes targeted interventions to efficiently test these hypotheses. Finally, the evidence gathered is used for knowledge **consolidation**, maintaining only those causal mechanisms that are invariant across different contexts and potentially useful for future tasks. This new framework allows the agent to curate what it learns, keeping only the knowledge that is causally valid and useful for future tasks. We formalize this approach in a three-level hierarchical architecture, detailed in Section 3, including high-level knowledge curation from mid-level tactical planning and low-level motor execution.

2 Related Work

World models for long-horizon control. Learned world models have become a cornerstone of modern model-based RL, enabling agents to plan efficiently in a latent space and significantly improve sample efficiency [17, 18, 19, 31]. By training on predictive objectives, these models learn dynamics that support imagination-based control. However, optimizing for prediction accuracy can cause models to overfit to spurious correlations within the training environment, failing to generalize under distributional shifts. Also, for long-horizon tasks, compounding prediction errors can quickly make imagined trajectories unreliable.

Hierarchical Reinforcement Learning (HRL) provides a solution, i.e., temporal abstraction or the options framework [39], allowing agents to reason with temporally chunked actions or high-level state representations (e.g., subgoals). Modern deep HRL methods learn multi-level policies for complex task decomposition [41, 25, 45]. While effective for structuring behavior within a given task, these HRL approaches do not explicitly focus on learning representations that are invariant across different contexts. This leaves a need for agents that can separate context-specific predictive knowledge required for immediate control from context-invariant causal knowledge required for robust generalization.

Agent learning to learn and explore. The challenge of cross-context generalization reframes agent learning as a meta-learning problem, where the agent must "learn to learn" by keeping learned

parameters or knowledge to future tasks [11, 42]. A key driver for this process is intrinsic motivation, which encourages skill acquisition without reliance on single extrinsic reward. Among various intrinsic objectives, empowerment—the mutual information between an agent's actions and future states—stands out as it quantifies an agent's capacity for control and influence over its environment [35]. This drive to discover controllable aspects of the world mirrors exploration patterns in human development, where interaction is often aimed at uncovering underlying causal structures [15]. However, applying empowerment naively over raw sensory states can lead to degenerate behaviors (e.g., the "noisy TV" problem) and does not guarantee the discovery of abstract, reusable concepts. The objective itself provides no mechanism for separating valuable causal knowledge from transient, environment-specific control. We propose to use empowerment not for undirected exploration, but as an intrinsic value function to evaluate and select causal knowledge based on its potential to expand the agent's capabilities across a distribution of imagined future tasks.

LLMs as priors for learning, reasoning, and curation. Recent work has explored Large Language Models (LLMs) as general prior for agents and world models, because they have vast stores of commonsense knowledge for planning, task decomposition, and reward specification [2, 23, 44, 5, 40, 26]. However, a fundamental limitation of these models is their lack of grounding in embodied experience; their knowledge is derived from statistical patterns in text, not from interaction with physical dynamics, making them prone to factual inaccuracies and causal fallacies [5, 40].

This motivates an 'LLM-in-the-loop' paradigm, where LLMs serve not as the central decision-maker but as a source of hypotheses validated through grounded interaction [43, 26]. Our framework lets LLMs propose diverse and plausible "imagined tasks," which are then uses to perform counterfactual tests on its causal models. This design bridges the gap between the broad priors of LLMs, the need for predictive control by world models and the causal knowledge an agent must acquire through interaction.

3 Framework: A Multi-World-Model Agent with Causal Curation

We propose an architecture that consists of two key ideas: (1) a *three-level hierarchy* that operates at different levels of representations of low-level control, mid-level planning, and high-level knowledge curation; and (2) a *dual-representation system* that learns predictive world models for in-context performance and causal models for across-context generalization. This design is inspired by the human ability to combine fast, intuitive motor control with slower, more deliberate causal reasoning.

3.1 Three-level hierarchy

We consider an agent in a continual learning setting, interacting with a sequence of contexts $c \sim \mathcal{D}$. Each context defines a distinct MDP, $\mathcal{M}_c = (\mathcal{S}, \mathcal{A}, P_c, R_c, \gamma)$ where the agent receives observations $o_t \in \mathcal{O}$ but does not directly observe context c. To succeed, the agent must adapt quickly and transfer knowledge effectively. Our architecture helps this by decomposing the agent's learning into three interconnected modules operating at different levels of abstraction. This framework readily extends to additional levels or continuous abstraction hierarchies.

Executor (Level 1) operates directly on raw and high-dimensional sensory inputs x_t^e (e.g., pixels, proprioceptive signals) and executes primitive actions a_t^e. It learns a context-specific predictive world model $\mathcal{W}_\theta^e : p_\theta^e\left(x_{t+1}^e \mid x_t^e, a_t^e\right)$. Executor ensures the high-fidelity dynamics necessary for precise motor control to achieve subgoals specified by the Controller.

Controller (Level 2) operates over abstract state space x_t^k using temporally extended actions, or options, $u_t \in \mathcal{U}$ [38]. Its predictive world model, $\mathcal{W}_\vartheta^k : p_\vartheta^k\left(x_{t+1}^k \mid x_t^k, u_t\right)$, operates at a lower temporal resolution, planning sequences of subgoals to accomplish tasks posed by the environment or the Curator. This level bridges high-level strategy and low-level execution.

Curator (Level 3) does not interact with the world directly. Instead it maintains a compact knowledge state x_t^q (e.g., a library $\mathcal{L}_t$ of context-invariant causal mechanisms and skills) [47]. Its world model, $\mathcal{W}_\zeta^q : p_\zeta^q\left(x_{t+1}^q \mid x_t^q, d_t\right)$, predicts how this *internal* knowledge evolves over learning given experienced tasks and imagined future tasks. The Curator's primary role is to decide which knowledge to acquire, test, and retain, guided by an intrinsic objective to maximize the agent's future capabilities.

3.2 Dual-representation heads

The agent's goal is specialization with fast adaptation to new contexts, thus we propose learning two complementary representations: (i) *context-specific predictive world models* for control and short-horizon planning, and (ii) a *context-invariant causal model* for counterfactual reasoning and robust generalization.

3.2.1 Context-wise learning of hierarchical world models

Within any given context, the Executor and Controller learn specialized predictive models. These models have complementary compression properties: the Executor's model, $\mathcal{W}^e_\theta$, is *time-compressed*, prioritizing high-fidelity predictions over short rollouts. In contrast, the Controller's model, $\mathcal{W}^k_\vartheta$, is *space-compressed*, operating on a low-dimensional abstraction of the state space, $x^k_t = \alpha(x^e_t, c)$, which allows for efficient, long-horizon planning with options.

These world models are trained with a standard probabilistic predictive loss (negative log-likelihood). For example, the objective function for Executor is defined as:

$$\mathcal{L}_{\text{executor}}(\theta) = \mathbb{E}_{\mathcal{D}}\big[-\log p^e_\theta\big(x^e_{t+1} \mid x^e_t, a^e_t\big)\big]\,. \tag{1}$$

To ensure these two hierarchical levels remain grounded and coherent, we enforce a consistency objective. An option u_t executed for Δt steps at the Executor level should result in a state transition that matches the one-step prediction made by the Controller in its abstract space. We can formalize this as a loss function:

$$\mathcal{L}_{\text{consistency}} = \mathbb{E}_{x^e_t, u_t \sim \mathcal{D}} \left\| \alpha(x^e_{t+\Delta t}, c) - \mathcal{W}^k_\vartheta(\alpha(x^e_t, c), u_t) \right\|^2 \tag{2}$$

where $x^e_{t+\Delta t}$ is the state reached after executing option u_t from x^e_t using the Executor's policy. Minimizing this loss forces the learned abstraction α and the Controller's world model $\mathcal{W}^k_\vartheta$ to form a consistent, multi-level representation of the environment's dynamics.

Intuitively, if the Controller predicts that the action *grasp* takes the gripper from "open, above object" to "closed, at object," then rolling out the Executor under that option should land, in the compact space, at the same place.

3.2.2 Across-context curation of causal models

In parallel, the Curator builds a single, context-invariant causal model. This model represents the agent's transferable knowledge about the world's underlying mechanisms. It is represented as a low-dimensional, factorized state $z_t = g(x_t)$ governed by a structural causal model (SCM), $z_{t+1} \leftarrow F(z_t, a_t, \epsilon_t)$, where actions are treated as interventions [32]. The mechanisms F are explicitly trained to be invariant across all experienced contexts, aiming for counterfactual reasoning $\text{do}(a_t), z_t)$ and transfering to novel context c.

Empowerment as an intrinsic value for knowledge. There is no ground-truth reward for "having the right theory." The Curator must therefore rely on an intrinsic objective. We argue that *empowerment* is a powerful candidate for this objective. Empowerment-driven exploration naturally leads agents to discover controllable aspects of their environment while maintaining behavioral diversity. This dual benefit is crucial: the search for controllability reveals causal relationships (interventions that reliably produce effects), while diversity ensures the agent discovers diverse causal paths and skills[35, 14, 47]. Correct causal knowledge, in turn, expands what the agent can reliably achieve, creating a virtuous cycle. Additionally, maintaining diverse capabilities through empowerment prepares the agent for unknown future tasks; it biases the agent toward discovering skills and mechanisms that are broadly useful. However, estimating this value is not straightforward. A practical approximation is to use LLMs as a semantic prior. We argue against directly prompting an LLM for the "value" of a causal hypothesis. While directly querying an LLM for causal hypothesis values would be ungrounded, we can use LLMs to generate diverse task scenarios. The agent then evaluates which causal knowledge would provide high empowerment across these scenarios through simulated rollouts (a form of preplay) [6].

Counterfactual reasoning on imagined tasks. While empowerment drives exploration, it alone doesn't ensure the agent learns transferable causal structure—it could simply memorize environment-specific controllers. Imagined future tasks address this by forcing the agent to reason counterfactually. This requires the agent to consider how discovered mechanisms would behave under different conditions, not just current ones. This is the central role of the accumulated causal knowledge (SCMs) in the library. The Curator fits the model parameters to be invariant across all visited contexts. Crucially, it enforces a *counterfactual consistency* loss that requires the SCM's predictions on imagined tasks to align with counterfactual rollouts performed in the context-specific world models of the lower levels. This ensures the causal model is deeply grounded in the agent's embodied experience.

The Curator's world model of knowledge. To manage this complex process, the Curator employs its own unique world model, $\mathcal{W}_\zeta^q : p_\zeta^q(x_{t+1}^q \mid x_t^q, d_t)$ to predict the evolution of the agent's knowledge. This is a *meta-level model* whose state, x_t^q, represents the agent's own knowledge—its library $\mathcal{L}_t = \{(z_i, F_i)\}_{i=1}^{|\mathcal{L}_t|}$ of skills z_i and associated causal models F_i. Its actions, d_t, are epistemic choices, such as which causal knowledge are preserved and discarded.

For example, given a candidate skill z_{new} with model $F_{z_{\text{new}}}$, we use three signals to determine adding it or not: 1) Empowerment of a skill z, $\mathrm{Emp}(z)$ (any consistent measure of control value); 2) Counterfactual validity of a causal model $\mathrm{CF}(F) \in [0, 1]$ (performance on held-out interventional/counterfactual queries); 3) Utility under the current task mixture $U_t(z)$ (expected performance uplift or predicted return). A candidate (z_{new} , F_{new}) is admitted iff:

$$\mathrm{Admit}\,(z_{\text{new}}\,, F_{\text{new}}\,) \Longleftrightarrow \begin{cases} \Delta\mathrm{Emp}_t\,(z_{\text{new}}\,) \geq \delta_{\text{add}}\,, \\ \mathrm{CF}\,(F_{\text{new}}\,) \geq \tau_{\mathrm{CF}}, \\ |\mathcal{L}_t| < K_{\max} \text{ or } U_t\,(z_{\text{new}}\,) > \min_{(z,F)\in\mathcal{L}_t} U_t(z). \end{cases} \tag{3}$$

Here $\delta_{\text{add}} > 0$ is a margin ensuring the new skill offers a meaningful marginal empowerment gain. $\tau_{\mathrm{CF}} \in (0, 1)$ is the required counterfactual validity threshold for causal soundness. If the library is full ($|\mathcal{L}_t| = K_{\max}$), the Curator evicts the lowest-utility item and inserts the candidate only when it strictly improves minimum utility. By planning in this "knowledge space," the Curator can select the epistemic actions expected to be most informative, guiding the agent on an efficient path toward a robust and transferable causal understanding of its world.

4 Discussion and Future Work

This work proposes a new perspective that *knowledge* is something an agent actively curates over the learning course, rather than a byproduct of learning a predictive world model with strong external supervision or intrinsic interactions. Specifically, we propose (i) separating time-compressed execution from space-compressed planning, (ii) scoring knowledge by its *empowerment* uplift on imagined tasks instead of asking a language model for ungrounded numeric values, and (iii) enforcing across-context, counterfactual consistency so that causal structure is portable. The outcome is not a single "big" model, but a well-curated library with options, goals, and causal mechanisms that the agent keeps only when they demonstrably increase what it can reliably make happen.

Inspired by human cognitive abilities, a general-purpose agent does not require perfectly identified causal graphs, nor do we assume the imagined tasks are "correct." The requirement is weaker: that the imagined scenarios are rich enough to expose when a hypothesis fails counterfactual tests, and that the learned structure improves performance on downstream tasks. Likewise, the proposed framework is compatible with standard model-based RL and does not hinge on a particular empowerment estimator.

This framework suggests several research directions worth exploring. Learning algorithms might benefit from explicitly combining empowerment maximization with causal invariance constraints, though the optimal balance remains an open question. Hierarchical curricula could potentially create productive loops where LLMs propose hypotheses for agents to test through intervention, though the effectiveness of such approaches needs empirical validation. Most importantly, evaluations should move beyond single-task benchmarks to assess genuine causal transfer, whether agents can apply discovered mechanisms in truly novel contexts.

References

[1] Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J Collins, Elizabeth Bourne, Gareth W Fearnley, Emma Tate, Holly X Smith, Larisa N Soldatova, et al. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of the Royal Society Interface*, 22(227):20240674, 2025.

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[3] Neil R Bramley, Peter Dayan, Thomas L Griffiths, and David A Lagnado. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, 124(3):301, 2017.

[4] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

[5] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*, pages 3676–3713. PMLR, 2023.

[6] Wilka Carvalho, Sam Hall-McMaster, Honglak Lee, and Samuel J Gershman. Preemptive solving of future problems: Multitask preplay in humans and machines. *arXiv preprint arXiv:2507.05561*, 2025.

[7] Christopher Cherniak. *Minimal rationality*. Mit Press, 1990.

[8] Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature human behaviour*, 8(10):1851–1863, 2024.

[9] Thomas Degris, Khurram Javed, Arsalan Sharifnassab, Yuxin Liu, and Richard Sutton. Step-size optimization for continual learning. *arXiv preprint arXiv:2401.17401*, 2024.

[10] Zhihong Deng, Jing Jiang, Guodong Long, and Chengqi Zhang. Causal reinforcement learning: A survey. *arXiv preprint arXiv:2307.01452*, 2023.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[12] Samuel J Gershman. 17 reinforcement learning and causal models. *The Oxford handbook of causal reasoning*, page 295, 2017.

[13] Mariel K Goddu and Alison Gopnik. The development of human causal learning and reasoning. *Nature Reviews Psychology*, 3(5):319–339, 2024.

[14] Alison Gopnik. Empowerment as causal learning, causal learning as empowerment: A bridge between bayesian causal hypothesis testing and reinforcement learning. 2024.

[15] Alison Gopnik and Henry M Wellman. Reconstructing constructivism: causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085, 2012.

[16] Tarun Gupta, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Marc Rigter, Ade Famoti, Ashley Juan Llorens, Jianfeng Gao, et al. The essential role of causality in foundation world models for embodied ai. *arXiv preprint arXiv:2402.06665*, 2024.

[17] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

[18] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[19] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[20] Khurram Javed and Richard S Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

[21] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

[22] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE, 2005.

[23] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. *arXiv preprint arXiv:2303.00001*, 2023.

[24] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[25] Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. *arXiv preprint arXiv:1712.00948*, 2017.

[26] Guillaume Levy, Cedric Colas, Pierre-Yves Oudeyer, Thomas Carta, and Clement Romac. Worldllm: Improving llms' world modeling using curiosity-driven theory-making. *arXiv preprint arXiv:2506.06725*, 2025.

[27] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.

[28] Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34:24111–24123, 2021.

[29] Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, et al. Exploring the limit of outcome reward for learning mathematical reasoning. *arXiv preprint arXiv:2502.06781*, 2025.

[30] Viraj Mehta et al. Filling the gaps: Llms for causal hypothesis generation. In *The Thirteenth International Conference on Learning Representations*, 2024.

[31] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.

[32] Judea Pearl. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450):428–431, 2000.

[33] Jonathan Richens, David Abel, Alexis Bellot, and Tom Everitt. General agents need world models. *arXiv preprint arXiv:2506.01622*, 2025.

[34] Jonathan Richens and Tom Everitt. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877*, 2024.

[35] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer, 2014.

[36] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[37] Steven A Sloman and David Lagnado. Causality in thought. *Annual review of psychology*, 66(1):223–247, 2015.

[38] Martin Stolle and Doina Precup. Learning options in reinforcement learning. In *International Symposium on abstraction, reformulation, and approximation*, pages 212–223. Springer, 2002.

[39] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[40] Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

[41] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International conference on machine learning*, pages 3540–3549. PMLR, 2017.

[42] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[43] Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, 2023.

[44] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. *arXiv preprint arXiv:2309.11489*, 2023.

[45] Mehdi Zadem, Sergio Mover, and Sao Mai Nguyen. Reconciling spatial and temporal abstractions for goal representation. *arXiv preprint arXiv:2401.09870*, 2024.

[46] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46, 2025.

[47] Hanqi Zhou, Fryderyk Mantiuk, David G Nagy, and Charley M Wu. Agent-centric learning: from external reward maximization to internal knowledge curation. *arXiv preprint arXiv:2507.22255*, 2025.

[48] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.

A Does Empowerment Lead to Causal Discovery?

We first analyze the conditions under which empowerment can drive causal discovery, and then argue why these conditions are too restrictive for realistic, open-ended learning.

A.1 Empowerment.

At its core, empowerment is the mutual information between an agent's actions A and the resulting future state S'. For a single state, it is the channel capacity:

$$\mathcal{E}(s) \;=\; \max_{P(a|s)} I(A; S' \mid S{=}s) \;=\; \max_{P(a|s)} \Big(H(S' \mid s) - H(S' \mid A, s) \Big), \tag{4}$$

where $A \sim P(a \mid s)$ and $S' \sim P(\cdot \mid s, A)$ [22, 35]. $\mathcal{E}(s) = \max_{p(a|s)} I(A; S' \mid S = s)$. We extend this to *macro-empowerment* for a sub-policy $\pi \in \Pi$ over a horizon T, defined as $\mathcal{E}_\pi(s) = I(A_{0:T-1}; S_T \mid S_0 = s, \pi)$. This measures the influence afforded by a specific skill.

A.2 Yes, in Principle, for Well-Defined Problems

Before detailing our framework, we first argue why, in a well-specified, toy and finite environment, an empowerment-maximizing agent is naturally driven to perform interventions that reveal causal structure. Consider an agent observing two correlated variables X and Y. Three scenarios are possible: (1) $X \to Y$ (X causes Y), (2) $Y \to X$ (Y causes X), or (3) $X \leftarrow Z \to Y$ (Common cause Z). An agent seeking to maximize its influence will inherently prefer actions that resolve this ambiguity.

Proposition 1 (Causal control advantage). Let A be the agent's action, X a controllable intermediate, and Y an outcome. If (i) X causally influences Y and (ii) X has at least one additional parent Z (i.e., $\exists Z : Z \to X$), then the empowerment under interventional policies strictly exceeds that under observational policies:

$$I(A; Y \mid \mathrm{do}(X)) \;>\; I(A; Y \mid X) \quad \text{iff} \quad X \to Y \text{ and } \exists Z : Z \to X. \tag{5}$$

Otherwise, the two quantities coincide.

Intuition. Conditioning on X leaves incoming edges into X intact, so dependence between A and Y can reflect confounding rather than genuine control. Intervening on X severs those incoming paths (e.g., from Z), isolating the effect of the agent's choice. This extra clarity about whether Y changes *because* of X yields strictly higher empowerment.

While Proposition 1 explains why empowerment drives intervention in a finite-variable setting, it doesn't explain how agents build reusable causal models. The key insight is that empowerment over longer horizons requires predicting which variables will remain controllable across contexts, exactly what causal models provide.

Proposition 2 (Causal invariance). Let $\mathcal{E}_\pi(s)$ denote the empowerment of policy π at state (or context) s. Across distributional shifts $s \sim P$, policies that act on causal parents exhibit more stable empowerment than policies that exploit merely correlational cues:

$$\mathrm{Var}_{s\sim P}\big[\mathcal{E}_{\pi_{\mathrm{causal}}}(s)\big] \;<\; \mathrm{Var}_{s\sim P}\big[\mathcal{E}_{\pi_{\mathrm{corr}}}(s)\big], \tag{6}$$

where π_{causal} manipulates causal parents of the outcome and π_{corr} relies on spurious correlations.

Intuition. Causal relations are invariant across shifts, so the effect of actions, and thus empowerment, remains stable. Correlational regularities can change with P, inflating the variability of empowerment. This stability encourages representations that separate causal from spurious factors, because only the former yield reliable empowerment across contexts.

A.3 Maybe Not in Realistic, Open-Ended Settings

While these propositions are encouraging, they rely on a pre-defined and typically small set of variables. In more realistic, open-ended worlds, vanilla empowerment fails for several key reasons:

1. **It operates on the wrong substrate.** Empowerment defined over low-level environment states (e.g., pixel values, joint angles) often leads to degenerate solutions like the "noisy

TV" problem. It encourages control over sensory data rather than meaningful aspects of the world, leading to overfitting to a specific environment's transition dynamics.

2. **It does not discover variables.** The propositions above assume the agent already knows about X, Y, and Z. A fundamental challenge is discovering these variables in the first place. Vanilla empowerment provides no direct pressure to abstract away from raw sensations to a vocabulary of causally meaningful entities and concepts.
3. **It is reactive, not counterfactual.** Empowerment is typically calculated based on an agent's present capabilities. It is not inherently forward-looking. For true generalization, an agent must reason counterfactually about what skills would be most useful for a *distribution of future tasks*, even those it has never seen.

These limitations reveal that raw empowerment is not enough. An agent needs a mechanism to discover abstract variables, ground its objectives in that abstract space, and curate its capabilities for future adaptation. Our framework is designed to address these exact challenges.