Assessing Adaptive World Models in Machines with *Novel Games*

Lance Ying[1,2], **Katherine M. Collins**[1,3], **Prafull Sharma**[1], **Cédric Colas**[1], **Kaiya Ivy Zhao**[1], **Adrian Weller**[3], **Zenna Tavares**[4], **Phillip Isola**[1], **Samuel J. Gershman**[2], **Jacob D. Andreas**[1], **Thomas L. Griffiths**[5], **François Chollet**[6], **Kelsey R. Allen**[7‡], **Joshua B. Tenenbaum**[1‡]

[1]MIT [2]Harvard University [3]University of Cambridge [4]Basis Research Institute [5] Princeton University [6] ARC Prize Foundation [7] University of British Columbia
[‡] Co-senior authors

Correspondence to `lanceying@mit.edu`

Abstract

Human intelligence exhibits a remarkable capacity for rapid adaptation and effective problem-solving in novel and unfamiliar contexts. We argue that this profound adaptability is fundamentally linked to the efficient construction and refinement of internal representations of the environment, commonly referred to as *world models*, and we refer to this adaptation mechanism as ***world model induction***. However, current understanding and evaluation of world models in artificial intelligence (AI) remains narrow, often focusing on static representations learned from training on massive corpora of data, instead of the efficiency and efficacy in learning these representations through interaction and exploration within a novel environment. In this Perspective, we provide a view of world model induction drawing on decades of research in cognitive science on how humans learn and adapt so efficiently; we then call for a new evaluation framework for assessing adaptive world models in AI. Concretely, we propose a new benchmarking paradigm based on suites of carefully designed games with genuine, deep and continually refreshing novelty in the underlying game structures — we refer to this class of games as ***novel games***. We detail key desiderata for constructing these games and propose appropriate metrics to explicitly challenge and evaluate the agent's ability for rapid world model induction. We hope that this new evaluation framework will inspire future evaluation efforts on world models in AI and provide a crucial step towards developing AI systems capable of human-like rapid adaptation and robust generalization — a critical component of artificial general intelligence.

1 Introduction

A hallmark of human intelligence is the capacity for rapid adaptation, solving new problems quickly under novel and unfamiliar conditions. Over evolutionary timescales, this adaptive intelligence has enabled humans to survive and flourish in a vast landscape of complex and ever-changing environments. In modern life, people are continually adapting to new social situations such as new laws, cultural environments, partners and foes—often with remarkable effectiveness and efficiency.

Decades of research in cognitive science suggests that a key mechanism supporting this rapid adaptation is the construction and refinement of mental models and intuitive theories to explain the world (Johnson-Laird, 1983; Gopnik and Wellman, 2012; Gelman and Legare, 2011; Tenenbaum et al., 2011; Gerstenberg and Tenenbaum, 2017; Ullman and Tenenbaum, 2020).

In the field of AI, these internal representations are often referred to as "world models" (Ha and Schmidhuber, 2018), an agent's representation of its environment, including objects, agents, and causal structures, which can be used to simulate and reason about the world. Building AI systems with more human-like world models and world-modeling capacities has been hypothesized as a crucial step towards building more general intelligent systems. The concept of world models has thus garnered significant recent interest in AI research, particularly regarding their structure, how they can be assessed, and whether today's AI systems truly possess them (Zhu et al., 2024; Ding et al., 2024; Hao et al., 2023; Andreas, 2024; Vafa et al., 2024).

However, despite increasing attention, the current ways that internal models are characterized and evaluated in AI systems often diverge importantly from the ways mental models have been studied in humans. Much existing evaluation focuses on static, low-level domain-specific representations learned from large, pre-collected datasets. In contrast, decades of cognitive science research highlights the ways human world models not only support rapid adaptation but are themselves highly adaptive. Our models are dynamically constructed and rapidly adjusted for new domains through active interaction, not merely learned offline from vast corpora. They operate across multiple scales of space, time and abstraction, with higher-level models constraining inferences and induction at lower levels and lower levels grounding the predictions of higher-level abstractions. In this Perspective, we refer to these capacities broadly as a capacity for ***world model induction***, which allows intelligent systems to quickly form and validate hypotheses about how new environments and tasks work, and use these hypotheses to guide action, exploration, and bootstrap further learning.

We expect that building and evaluating AI systems capable of this kind of rapid world model induction will be critical for achieving robust, general AI capable of functioning effectively in the complex and fast-changing real world, and especially in *human* worlds – the environments that human beings have evolved in, created, and are continually changing and re-creating.

To drive AI progress towards this goal, and to be able to measure that progress, we argue for a comprehensive evaluation framework grounded in the cognitive science theories and experimental paradigms that have been used to study world model induction in humans. We propose that games are a uniquely advantageous domain for evaluating these capabilities in AI, given their inherently rich, often hierarchical structures in concepts and skills and their well-controlled environments. Concretely, we introduce an evaluation paradigm centered around the concept of *novel games*. Within this framework, a *novel game* is defined not simply by unseen instances or parametric variations within a familiar game structure, but by environments with structured novelty, where underlying rules, mechanics, object properties, or objectives are initially unknown, hidden, or dynamically changing. Success requires agents to rapidly infer these latent dynamics and causal structures through active, limited interaction and exploration, effectively performing world model induction on-the-fly.

We hope this Perspective will guide future evaluative work on AI world models and thereby driving progress towards machines that can efficiently learn, generalize, and adapt with human-like flexibility and robustness in complex, dynamic real-world environments.

2 World Models in Humans and Machines

The precise definition of world models is often debated (Ding et al., 2024). For the purposes of this paper, we define *world model* as an agent's internal representation of an environment, including its dynamics, rules, objects, and underlying causal relations. The utility of such a model lies in its ability to allow agents to efficiently simulate different world states for effective decision making, planning, and problem solving.

The study and evaluation of representations within today's AI systems has garnered significant interest in recent years. However, much of the existing work on AI world models often characterizes and evaluates these representations in a relatively narrow scope, frequently treating them as static representations primarily capturing low-level domain features learned from large, pre-collected datasets, such as whether a model can predict the next frame of a video or future states (LeCun, 2022; Ha and Schmidhuber, 2018), or recover a visual or spatial representation of the environment (Vafa et al., 2024; Li et al., 2023). Although these offer valuable insights into the inner representations, we contend that this understanding and evaluation is insufficient for developing AI systems capable of human-level sample efficient learning and adaptation in the open world.

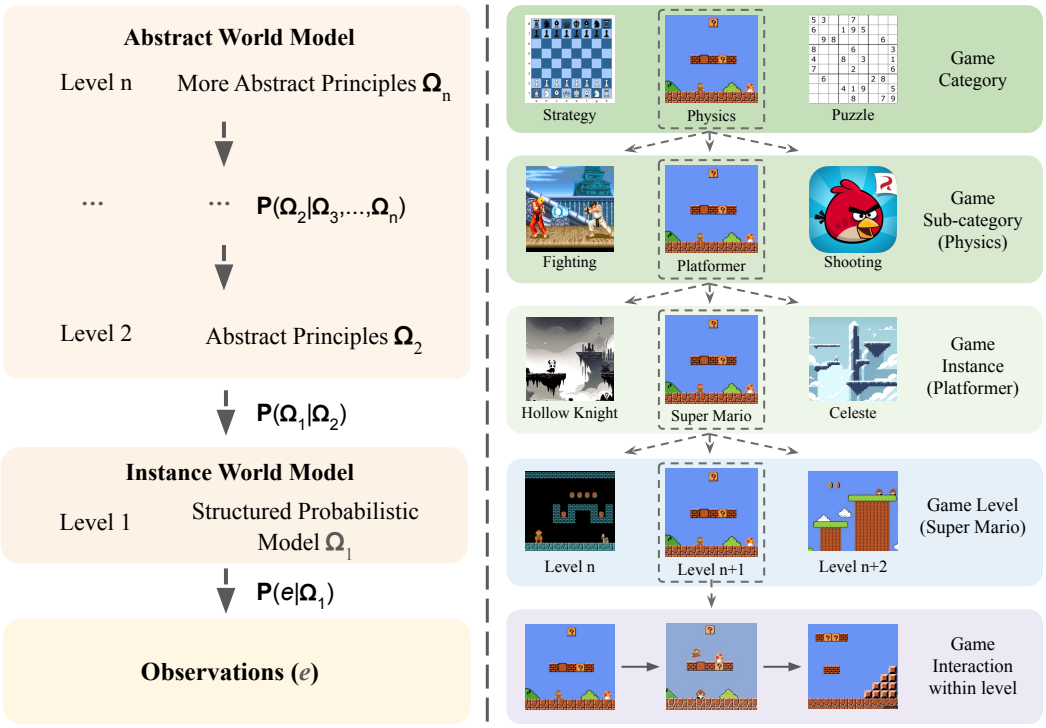

Figure 1: **Framework for characterizing world models across different levels of abstractions. a)**. World models within a hierarchical Bayesian framework. The structured probabilistic model Ω_1 (ad-hoc world model) generates expectations about possible observations e, while abstract knowledge and principles (abstract world model) $\Omega_2, \Omega_3, \ldots$ generate the space of possible structures for Ω_1. Each level of abstraction generates hypotheses and probability distributions that support learning at the level below. Figure adapted from Tenenbaum et al. (2006). **b)**. The hierarchical structure of games is analogous to many aspects of the human world model hierarchy. The world model learned at each game level can be ad-hoc and specific. On the other hand, meta-learning enables agents to learn domain-general principles at higher levels of abstraction.

For adaptation to complex and changing environments, an agent's world models cannot be static representations learned once and fixed. The world is inherently dynamic and often unpredictable. Agents frequently encounter novel situations where previously learned knowledge may be incomplete, only partially applicable, or even become obsolete, necessitating continuous refinement and potentially significant restructuring of the agent's mental model of the environment. Therefore, an adaptive agent's world model must be capable of being dynamically updated and adapted in response to new experience. This continuous process of inferring and revising the world model through interaction is what we refer to as **world model induction**.

A key characteristic of human world model induction is its **sample efficiency**. How do humans achieve such remarkable sample efficiency in learning about the world? Extensive research in cognitive science has formulated human learning and intuitive theory induction within a hierarchical Bayesian framework (Gopnik and Wellman, 2012; Gelman and Legare, 2011; Tenenbaum et al., 2011; Ullman and Tenenbaum, 2020), where theories and concepts are learned and represented at different levels of abstraction. We can draw a similar conceptual framework for world models, as shown in Figure 1.

In our framework, we partition an agent's world model, denoted as Ω, into two categories of representation, distinguishing between an *instance world model* and *abstract world models*. An instance world model Ω_1, at the lowest level of abstraction, is often a detailed, structured, and domain-specific representation pertaining to a specific instance, which can be constructed on-the-fly to explain observations within that environment, for example, a cognitive map of New York City. Abstract world

models $\Omega_2, \Omega_3, \ldots$ are more abstract generalizable concepts and principles applicable across domains. For example, one's understanding of real-world physics can be applied even if one moves to a new city. This hierarchical world model structure is key to human adaptation as abstract world models can provide informative priors for a sample-efficient construction of ad-hoc instance world models for interacting and problem-solving within new domains.

Learning can be understood as the construction and refinement of such hierarchical world model. When an agent receives new observations (e), its beliefs about the underlying world model Ω are updated. The posterior probability distribution over possible world models, given new data e, is computed by inverting the generative model:

$$\begin{aligned} P(\Omega|e) &\propto P(e|\Omega)P(\Omega) \\ &\propto P(e|\Omega_1)P(\Omega_1|\Omega_2)\ldots P(\Omega_n|\Omega_{n+1}) \end{aligned} \tag{1}$$

Here, the likelihood term $P(e|\Omega)$ represents the probability of observing the data e given a specific world model Ω, while the prior $P(\Omega)$ represents the agent's beliefs about Ω before the new observation.

The efficacy and sample efficiency of this Bayesian update process are significantly enhanced when the agent is not merely a passive observer but actively seeks out informative data. Insights from cognitive science and developmental psychology, particularly the "the child as scientist" framework (Schulz, 2012; Gopnik and Wellman, 1992; Gopnik, 1996), suggest that human learning is characterized by hypothesis-driven exploration. Starting at infancy, human learning involves actively planning and designing "experiments" — actions that (intentionally or not) effectively generate informative observations that can discriminate between competing hypotheses about the underlying world model Ω. We expect that this kind of active hypothesis-driven approach to generating data is critical for agents to converge on an accurate representation of a new, unfamiliar domain with minimal interaction and observation, and more generally for sample-efficient world model induction.

Adaptation via World Model Induction

Human-like adaptive intelligence necessitates world model induction at different levels of abstraction. This affords agents a number of core behavioral capabilities that are crucial for success in complex and fast-evolving environments. These include:

1. **Rapid Learning in New Domains:** The ability to achieve proficiency quickly in a wide range of previously unseen domains. This learning is facilitated by a combination of mechanisms, including generalization from sparse experiences, efficient goal-directed exploration, and the intelligent use of data sources beyond direct trial-and-error interaction (e.g., information derived from language or social observation).
2. **Robust Generalization within a Domain:** The capacity to generalize effectively to new and varied situations encountered *within* a newly learned domain. This includes adapting flexibly to novel perceptual inputs, understanding the behavior and affordances of previously unseen object types, handling modified consequences for actions, or pursuing altered goals within that domain's structure.
3. **Cross-Domain Generalization and Meta-Learning:** The development of meta-learning capabilities, enabling faster and more efficient adaptation to new domains by leveraging prior world models. This reflects the human ability to build generalizable knowledge (e.g., intuitive physics, intuitive psychology) that can bootstrap learning and generalize broadly to new tasks, even those with fundamentally different domain characteristics (Spelke and Kinzler, 2007; Allen et al., 2020; Lake et al., 2017; Chollet, 2019).

The capacity for rapid world model induction and the associated adaptive capabilities outlined above will be crucial for a wide range of practical applications that AI designers may target, such as adapting to new work environments and collaborating effectively with new human or artificial partners, especially when new tools or protocols are introduced. These capacities are also crucial for AI systems intended to function as AI scientists (Wang et al., 2023; Bengio et al., 2025; Geng et al., 2025), as human scientific discovery fundamentally involves actively forming hypotheses about the world, at different levels of abstraction, and designing experiments to validate these provisional models, mirroring the process of hierarchical world model induction (Henderson et al., 2010).

Despite the critical importance of world model induction for achieving human-like intelligence, there is a lack of comprehensive evaluation frameworks specifically designed for such world models in AI systems. Existing evaluations of world models in AI models often focus on assessing static world models learned from large, pre-collected datasets and extensive offline training (Vafa et al., 2024; Li et al., 2023), rather than measuring the efficiency and flexibility of models in learning and adapting world models through online exploration and interaction in genuinely novel domains.

In this perspective, we call for future AI evaluation efforts to holistically assess world models in machines according to the framework outlined above. We contend that games provide particularly rich and controlled environments uniquely well-suited for systematically evaluating rapid model adaptation and the process of world model induction. The remainder of this paper details how games can serve this purpose and proposes a new game-based evaluation in subsequent sections.

3 Games as a Benchmark for Intelligence

Games are universal cultural artifacts and have been commonly used as a measure of intelligence (Cleveland, 1907). While the definition of games is frequently debated, in this paper, we follow previous work on using games to study intelligence (Allen et al., 2024) and define games as "facilitators that structure player behavior and whose main purpose is enjoyment" (Newell et al., 1972).

Games have long served as valuable environments for studying machine intelligence by the AI community (Campbell et al., 2002; van Opheusden et al., 2023; Silver et al., 2016; Yannakakis and Togelius, 2018; Vinyals et al., 2019; Shannon, 1950; Newell, 1955; Chase and Simon, 1973). They strike a unique balance by offering clear rules, goals, and feedback while also requiring agents to engage in complex planning, learning, and abstraction. This combination of formal structure and behavioral complexity makes them especially well-suited for probing how intelligent systems — biological or artificial — make decisions under uncertainty. Formally, many games can be modeled as Partially Observable Markov Decision Processes (POMDPs), which define a task in terms of hidden states, observations, transitions, and rewards (Kaelbling et al., 1998).

Many established AI benchmarks, particularly those involving complex games (e.g. Atari (Bellemare et al., 2013), Go (Silver et al., 2016), StarCraft (Vinyals et al., 2019)) for reinforcement learning agents, follow a training/testing paradigm in which agents are optimized over millions or billions of interaction steps. While such systems can achieve superhuman performance, their success typically reflects extensive optimization within fixed environments rather than rapid, human-like adaptation to new environments.

To address the limitations of evaluating AI systems solely within the distribution of their training data, various reinforcement learning (RL) benchmarks have introduced forms of task variation to test generalization capabilities. However, much of the generalization evaluation in RL has focused primarily on measuring changes in raw task performance rather than **evaluating the *process* of adaptation itself**. There is often insufficient focus on whether these tasks truly necessitate the synthesis of new world models, how efficiently an AI model actually constructs these internal representations, or how these models *evolve* over time through interaction with the novel environment.

These limitations collectively highlight a critical gap in current AI evaluation paradigms. While they succeed in measuring performance under varying conditions or with some task generalization, they fall short of assessing the core human capacity to actively construct and dynamically adapt internal world models based on limited online experience and interaction. This ability demands an evaluation framework specifically tailored to reveal an agent's model-building capabilities. To address this need, we next introduce an evaluation paradigm centered on the use of *novel games* and detail how the design of these games, alongside appropriate metrics, can provide a robust method for assessing rapid world model induction in AI.

4 Assessing Adaptive World Modeling in AI with *Novel Games*

In this section, we propose an evaluation paradigm based on a class of games we call ***novel games*** for assessing the capacity for adaptive world modeling in AI. We are using the phrase *novel games* to refer to **games with genuine, deep, and continually refreshing novelty in the structure of the environments and goals** presented to the players. These games require players to construct new world models or modify their models when first learning the game and dynamically throughout their

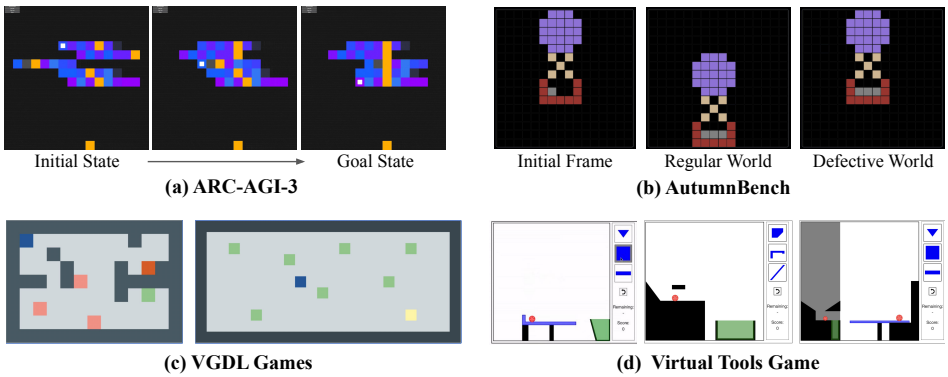

Figure 2. **Case studies of novel environments for testing world model induction**

a) ARC-AGI-3 (ARC Prize, 2025) is an interactive reasoning benchmark. Players are not given instructions about the gameplay or win conditions; instead, they must infer game rules and objectives through interaction.

b) AutumnBench (Basis, 2025) evaluates an agent's ability to discover latent mechanics through interactive exploration of grid-world environments. The evaluation follows a two-phase protocol: interaction and test. During interaction, agents explore environments freely without rewards or goals. The subsequent test phase evaluates their understanding through three tasks: masked frame prediction, defect detection, and planning. In the featured example, through interaction the agent is expected to discover the rule that adding a grey block lowers the ballon, which makes the right-most frame an anomaly.

c) VGDL games (Schaul, 2013; Perez-Liebana et al., 2019) were used by Tsividis et al. (2021) to evaluate an agent's capability to discover rules and objectives in ambiguous environments when no language instructions are provided. In their games, the agent can press a few keys on a keyboard to explore the environment to pass each level. Humans generally learn to play these new games in a matter of minutes.

d) Virtual Tools Game (Allen et al., 2020) tests rapid learning in physical reasoning scenarios. The goal is to select and drop one of the tools on the right so that the red ball ends up in the green bin. Allen et al. (2020) show that humans quickly solve these tasks by leveraging their intuitive understanding of physics. This showcases robust cross-domain generalization.

play, across boards, screens or levels. In the following sections, we first discuss the desiderata for such games, and then we propose a set of metrics for evaluating AI systems within them.

4.1 Desiderata for Designing *Novel Games*

The key requirement for our AI evaluation paradigm is the inherent novelty in the underlying game structure. These games must offer genuinely new adaptation challenges, meaning they are significantly distinct from established and widely studied games (like Chess, Go, or classic Atari titles) and necessitate new world models. This distinctiveness is crucial to prevent AI systems from reusing existing world models to solve the task or exploiting readily available online resources such as wikis and walkthroughs that describe optimal strategies for existing popular games.

However, the space of all possible *novel games* is infinitely large as one can construct new games with any arbitrary mechanics. We propose *novel games* should be grounded in the kinds of diverse, highly dynamic and novel environments encountered by humans, thus providing a testbed for how well an AI system can learn and adapt in all kinds of (game) worlds intuitive to humans, either alone or with humans.

4.1.1 Desideratum 1: Novelty in Game Structures

In this section, we articulate key features for the design of domains with genuine novelty in their underlying structures for testing the three aspects of adaptive capabilities listed in Section 2:

Rapid Learning and Theory-driven Exploration We encourage the design of game environments where the underlying mechanics are not fully transparent or pre-specified to the player. Instead, crucial aspects – including types of objects, specific rules governing interactions, affordances and properties of objects, or the consequences of actions – should be **partially or entirely latent**, requiring the agent to infer them through active gameplay, exploration, and experimentation. We highlight three such game environments from previous studies in Figure 2.

This design compels the AI agent to function as an active learning system, dynamically constructing an understanding of its environment. This exploratory process should ideally reflect aspects of theory-driven learning (Ullman and Tenenbaum, 2020; Tsividis et al., 2021). The agent must be capable of forming hypotheses about latent rules or object behaviors based on observation, strategically planning and executing 'experiments' through its actions to test these hypotheses, and refining its internal model based on the observed outcomes. Effectively pursuing this form of active, model-building learning necessitates setting *epistemic goals*—objectives focused on acquiring knowledge and reducing uncertainty about the game's state and mechanics.

Robust Generalization within a Domain Effective adaptation within a learned novel domain requires robustness to variations and changes occurring within that specific environment's structure. *Novel games* should be designed to feature multiple levels or configurations, introduce new object types with distinct properties, modify existing rules or mechanics, or alter goals and outcomes over the course of interaction. Crucially, these games can also incorporate mechanics that cause the environment, including its rules and dynamics, to **evolve dynamically over time**, potentially influenced by the player's actions or external events. This dynamic aspect necessitates that the agent continuously monitors the environment, detects changes, and updates its internal world model online.

Flexible Generalization across Domains To evaluate the capacity for cross-domain generalization and meta-learning, the benchmark should include game sets where **abstract principles or models learned can be productively transferred to a new game**, despite significant differences in surface rules or mechanics. For example, training on games involving various scenarios governed by a consistent set of physics rules (e.g., gravity, momentum) allows an agent to induce an abstract "intuitive physics" model. This model can then be transferred to accelerate adaptation in a new game featuring new objects and tasks but operating under similar physical laws, enabling the agent to predict outcomes more effectively from the outset.

4.1.2 Desideratum 2: Intuitive and Learnable for Human Players

For *novel games* to stress-test AI models' capability to adapt in the human world, their core mechanics and objectives should be fundamentally **intuitive and learnable for average human players**. This criterion is essential because a key goal of this evaluation paradigm is to measure human-like adaptation skills. Games that humans find intuitive are likely structured in ways that resonate with fundamental human inductive biases: the inherent cognitive predispositions and learning mechanisms shaped by the cultural and physical environment humans inhabit (Allen et al., 2024; Dubey et al., 2018). Ensuring human learnability serves practical purposes: it allows for benchmarking AI performance directly against human capabilities, provides a valuable constraint on the complexity and potential arbitrariness of the game generation process, and aligns the evaluation with the broader goal of developing AI that can learn from and collaborate with humans in novel scenarios.

4.1.3 Desideratum 3: Diversity in World Models and Learning Mechanisms

To span the diverse array of challenges people—and AI systems in a human-world may face—the benchmark game suite should encompass significant **diversity** to necessitate different *types of world models* that agents are compelled to induce. For example, while some games may primarily involve learning about spatial relationships or object physics, others can require understanding and modeling other agents in multi-agent games (whether competitive or collaborative with other human or artificial agents). Successfully adapting in multi-agent scenarios often requires agents to develop sophisticated mental models about other agents, inferring and representing their goals, beliefs, intentions, or emotional states (often referred to as Theory of Mind (Gopnik and Wellman, 1992)), which constitutes a crucial aspect of human social adaptation.

Furthermore, the benchmark should feature diversity in the **learning mechanisms** available to the agent. Some games can be designed to offer minimal or no explicit instructions, compelling the agent to induce the world model predominantly through interaction and exploration. Conversely, other games could provide structured linguistic instructions, demonstrations, or tutorials, allowing for the evaluation of how well agents can leverage external, often multimodal, information to accelerate model construction. Incorporating scenarios where information about mechanics or objectives is conveyed implicitly through social means, such as Non-Player Characters (NPCs) that demonstrate actions or use language, can provide a critical way to test learning through observation and social scaffolding like people.

4.1.4 Combining Desiderata in a Generative Framework

A central challenge in this evaluation paradigm is the continual provision of games that rigorously satisfy our desiderata. Specifically, the inherent novelty of these games is ephemeral; as AI systems (and indeed, humans) gain experience with their mechanics, the games quickly become familiar, thereby undermining their utility as tests of adaptation to truly novel situations.

We propose that game benchmarks should be thought of as a **generative process** over such games which can continually sample new *novel games* that satisfy our desiderata. Like language, games can be compositional and continually reconfigured. Modifications can vary the game mechanics, partners, and other game features (see Figure 3). This would allow the game benchmark to continue to evolve and cover a large space of novel and diverse environments that AI would need to adapt to, thus mitigating overfitting.

4.2 Evaluation of AI Agent's World Modeling Capacity

Once we have designed games that pose meaningful challenges on adaptive world modeling for AI, we need a comprehensive evaluation framework to characterize the internal world models learned by the agent.

Sample Efficiency in Adaptation A measure of learning efficiency is how quickly a model can achieve proficiency with limited experience. One could evaluate this by providing a restricted "budget" of training attempts (trials) within a game level and assessing performance. This budget can be varied to provide a fine-grained understanding of learning dynamics and adaptability under different constraints. For example, one can measure performance after a fixed number of attempts, or quantify the number of game-plays required to reach average human performance (e.g. Lake et al. 2017).

Qualitative Analysis of Exploration and Learning Behavior Beyond quantitative metrics, a qualitative analysis of how agents explore and learn within novel environments can reveal crucial insights into their capabilities for world model induction. Different learning approaches often manifest in distinct exploration patterns, reflecting their strategies for gathering information and inferring rules. For instance, Tsividis et al. (2021) finds that human players tend to exhibit targeted and efficient exploration when learning to play novel games, focusing on areas relevant to understanding the game mechanics, while other learning algorithms such as DDQN (Van Hasselt et al., 2016) often displays highly diffuse and less directed exploration, indicative of a struggle to efficiently form coherent internal representations of the environment.

Probing Internal World Models To properly assess rapid world model induction, it is essential to gain insight into the nature of the internal representations that the agent constructs and how these representations are dynamically updated in response to new observations and actions.

The methods for inspecting these internal world models are heavily dependent on the agent's architecture. For models based on explicit program synthesis or symbolic reasoning, the inferred world model may be directly interpretable as the synthesized program or set of rules (Tsividis et al., 2021; Das et al., 2023). This offers a transparent view of the agent's current representation of the environment, and this allows direct comparison to the kinds of hierarchical representations in humans.

For neural networks, inspecting the internal world model is more complex, typically involving analyzing representation spaces, activation patterns, attention mechanisms, or tracing reasoning processes through the network (Vafa et al., 2024). Techniques such as probing specific network layers

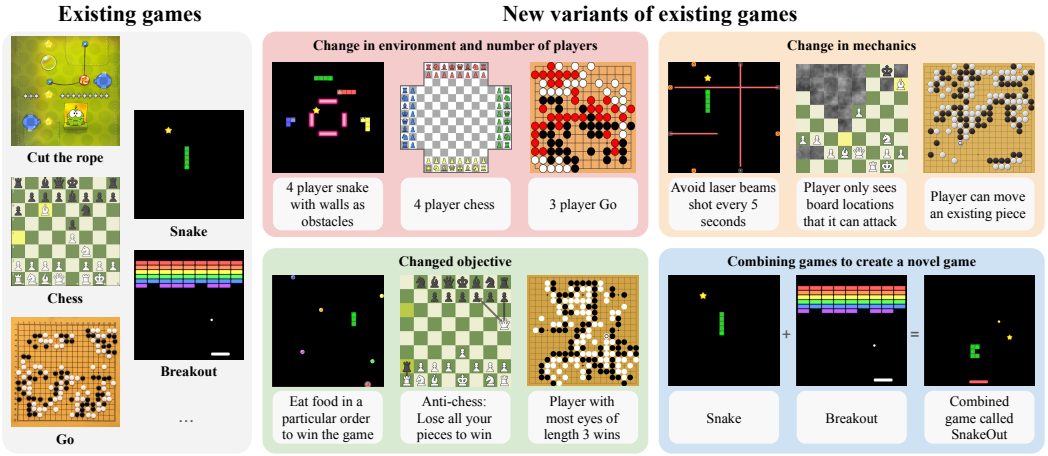

Figure 3: **Creating *novel games*.** There are many ways that researchers can create *novel games*. Variants could be sourced from existing games by modifying the environment and number of players, mechanics, or objective function. *Novel games* could also be formed by combining existing games.

for learned features related to game mechanics or dynamics can reveal aspects of the implicit world model. For many of the large foundation models that can interact with humans in natural language, we can also examine their understanding of the game mechanics through targeted question answering at different levels of abstraction.

By probing these internal representations and their changes over time as the agent interacts with a game environment, we can gain crucial qualitative insights into how the AI agent is actively inferring, representing, and revising its understanding of the world.

5 Discussion and Looking Forward

In this Perspective, we have argued that a critical component for developing truly general and robust artificial intelligence lies in its capacity for adaptation to novel circumstances. This adaptive capability is fundamentally linked to the agent's ability to rapidly induce and dynamically refine internal world models when confronted with unknown environments. We then introduce an evaluation paradigm centered around carefully constructed *novel games*. This framework is specifically designed to evaluate AI systems on their capacity for adaptive world modeling, which is essential for efficient learning and robust generalization in dynamic, unforeseen environments where underlying rules and structures are often hidden from the agent.

While we believe this paradigm offers a valuable path forward, we acknowledge several important outstanding questions and challenges that warrant future investigation and refinement. A fundamental question that may arise regarding our central thesis is the extent to which hierarchical and adaptive world models are truly necessary for rapid and efficient adaptation. Our Perspective is strongly motivated by and aligned with extensive evidence from cognitive science and developmental psychology (Tenenbaum et al., 2011; Gopnik and Wellman, 2012), which highlights the crucial role of constructing and refining internal models in human learning and adaptation. The extent to which such model-free approaches can achieve human-level rapid, sample-efficient adaptation and robust generalization remains an open empirical question.

Second, while we can evaluate performance on the tasks within *novel games*, directly measuring the quality and efficiency of internal world model induction at different levels of hierarchy presents its own challenges. Developing metrics that specifically quantify how well an agent has inferred the latent rules or dynamics—beyond just task success—and how efficiently it updates this understanding over time is crucial. This may involve developing probing techniques, counterfactual evaluation methods based on the inferred model, or analyzing the structure of internal representations.

6 Conclusion

In this paper, we have argued that building human-like adaptability in machines necessitates adaptive world models, which affords sample efficient world model induction in any new domains. We then proposed a novel framework for assessing adaptive world models centered on the concept of *novel games*. We believe this proposed evaluation paradigm holds significant potential to serve as a core component in assessing current AI models and drive research towards systems that exhibit the rapid, flexible, and robust adaptability characteristic of human intelligence, thus contributing meaningfully to the ambitious pursuit of artificial general intelligence.

References

Allen, K., Brändle, F., Botvinick, M., Fan, J. E., Gershman, S. J., Gopnik, A., Griffiths, T. L., Hartshorne, J. K., Hauser, T. U., Ho, M. K., et al. (2024). Using games to understand the mind. *Nature Human Behaviour*, pages 1–9.

Allen, K. R., Smith, K. A., and Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310.

Andreas, J. (2024). Language models, world models, and human model-building.

ARC Prize (2025). Arc-agi-3. Retrieved from https://arcprize.org/arc-agi/3.

Basis (2025). Autumnbench: World model learning in humans and ai.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. *Journal of artificial intelligence research*, 47:253–279.

Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., et al. (2025). Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*.

Campbell, M., Hoane Jr, A. J., and Hsu, F.-h. (2002). Deep blue. *Artificial intelligence*, 134(1-2):57–83.

Chase, W. G. and Simon, H. A. (1973). The mind's eye in chess. In *Visual information processing*, pages 215–281. Elsevier.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Cleveland, A. A. (1907). The psychology of chess and of learning to play it. *The American Journal of Psychology*, 18(3):269–308.

Das, R., Tenenbaum, J. B., Solar-Lezama, A., and Tavares, Z. (2023). Combining functional and automata synthesis to discover causal reactive programs. *Proceedings of the ACM on Programming Languages*, 7(POPL):1628–1658.

Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al. (2024). Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*.

Dubey, R., Agrawal, P., Pathak, D., Griffiths, T. L., and Efros, A. A. (2018). Investigating human priors for playing video games.

Gelman, S. A. and Legare, C. H. (2011). Concepts and folk theories. *Annual review of anthropology*, 40(1):379–398.

Geng, J., Chen, H., Arumugam, D., and Griffiths, T. L. (2025). Are large language models reliable ai scientists? assessing reverse-engineering of black-box systems.

Gerstenberg, T. and Tenenbaum, J. B. (2017). Intuitive theories.

Gopnik, A. (1996). The scientist as child. *Philosophy of science*, 63(4):485–514.

Gopnik, A. and Wellman, H. M. (1992). Why the child's theory of mind really is a theory.

Gopnik, A. and Wellman, H. M. (2012). Reconstructing constructivism: causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085.

Ha, D. and Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. (2023). Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Henderson, L., Goodman, N. D., Tenenbaum, J. B., and Woodward, J. F. (2010). The structure and dynamics of scientific theories: A hierarchical bayesian perspective. *Philosophy of Science*, 77(2):172–200.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*.

Newell, A. (1955). The chess machine: an example of dealing with a complex task by adaptation. In *Proceedings of the March 1-3, 1955, western joint computer conference*, pages 101–108.

Newell, A., Simon, H. A., et al. (1972). *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ.

Perez-Liebana, D., Liu, J., Khalifa, A., Gaina, R. D., Togelius, J., and Lucas, S. M. (2019). General video game ai: A multitrack framework for evaluating agents, games, and content generation algorithms. *IEEE Transactions on Games*, 11(3):195–214.

Schaul, T. (2013). A video game description language for model-based or interactive learning. In *2013 IEEE Conference on Computational Inteligence in Games (CIG)*, pages 1–8. IEEE.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in cognitive sciences*, 16(7):382–389.

Shannon, C. E. (1950). Xxii. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(314):256–275.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Spelke, E. S. and Kinzler, K. D. (2007). Core knowledge. *Developmental science*, 10(1):89–96.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

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

Ullman, T. D. and Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2(1):533–558.

Vafa, K., Chen, J., Rambachan, A., Kleinberg, J., and Mullainathan, S. (2024). Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975.

Van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., and Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, pages 1–6.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Yannakakis, G. N. and Togelius, J. (2018). *Artificial intelligence and games*, volume 2. Springer.

Zhu, Z., Wang, X., Zhao, W., Min, C., Deng, N., Dou, M., Wang, Y., Shi, B., Wang, K., Zhang, C., et al. (2024). Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*.