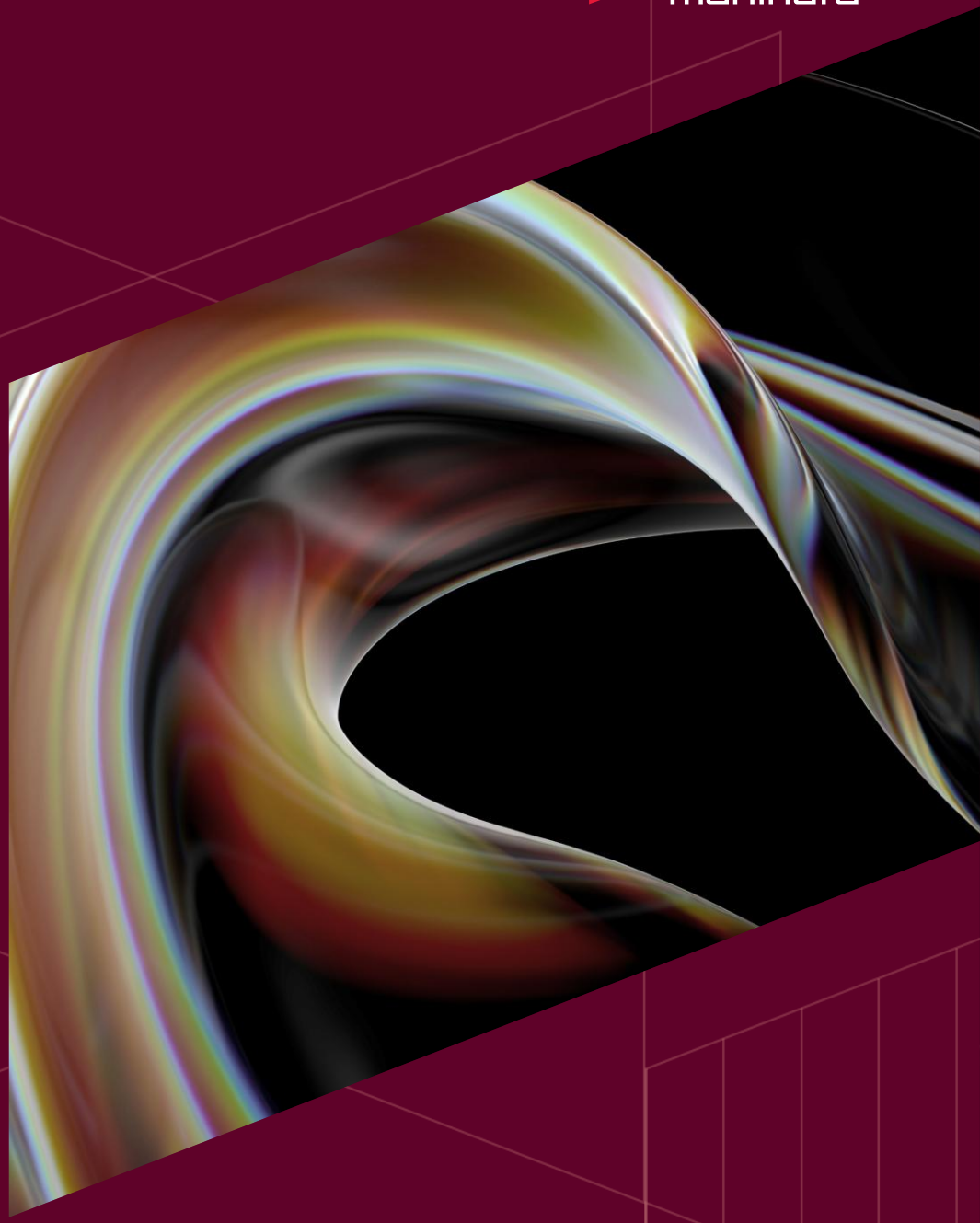
TECH mahindra

Whitepaper

Building Adaptive AI with Symbolic Deep Reinforcement Learning



Author

Nikhil Malhotra

Executive Summary

The world of AI has so far been limited by its extensive reliance on LLMs. Undoubtedly, an LLM's transformer architecture has played a significant role in enabling it to predict the next token, but such models are highly fragile and autoregressive. Compared to its predecessors, the model establishes an attention mechanism across word categories within a sentence, enabling it to disambiguate words more effectively. On top of that, the model has convinced users that it demonstrates greater nuance and can respond in the user's native language. In reality, it just statistically predicts the next word. The claims that the model has gained greater intelligence are baseless, too. Especially because the model does not fully comprehend the world it exists in and doesn't understand the causal relationship between various objects, tasks, and actions. In this approach, causality suffers, and so does general intelligence.

This whitepaper discerns the key mechanisms in how humans learn and presents a roadmap on how AI can replicate them. It also recommends SDRL as a framework that embeds adaptive intelligence into models rather than relying on statistical pattern matching.

Table of Contents

1) Introduction 04

2) The World Model Formulation 05

3) The Opportunity: How Can AI Learn 07

4) Advent of Symbolic Deep Reinforcement Learning 09

5) Symbolic Priors: Embedding PreBEK in SDRL 11

6) Illustrative Example 13

7) The Final Word 14

Introduction

The limitations of LLMs in their learning endeavors present a stark reality: transformer-architecture-based AI models merely simulate linguistic competence and are disconnected from the world they operate in. This is not real intelligence, and these models majorly lack a genuine understanding of the world. Human intelligence, on the other hand, offers a compelling alternative blueprint for AI training. Here's a rundown on how humans understand the world.

According to neuroscience and cognitive science, neuro-symbolism plays a significant role in how we learn and perceive the world around us.[1] Additionally, humans possess the speed and power with which we can learn and make decisions. These inherent capabilities together help us better understand the world.

This hybrid neuro-symbolic nature of cognition is most evident in the neocortex, which forms the intelligent brain of Homo Sapiens, rather than the Medulla Oblongata or the midbrain, which primarily governs fight-or-flight responses. The reflective AI training blueprint, therefore, focuses on neocortex principles that mirror fast, adaptive, and grounded intelligence.

Furthermore, from infancy, humans exhibit unique characteristics, such as inherent pattern recognition and genetically encoded primitive knowledge. This pre-learned knowledge is known as pre-birth encoded knowledge (PreBEK). PreBEK plays an active role in the development of new memories and relationships throughout a human's life. This foundational symbolism gets substantiated by post-birth encoded knowledge (PoBEK). Together, they guide human understanding and behavior.

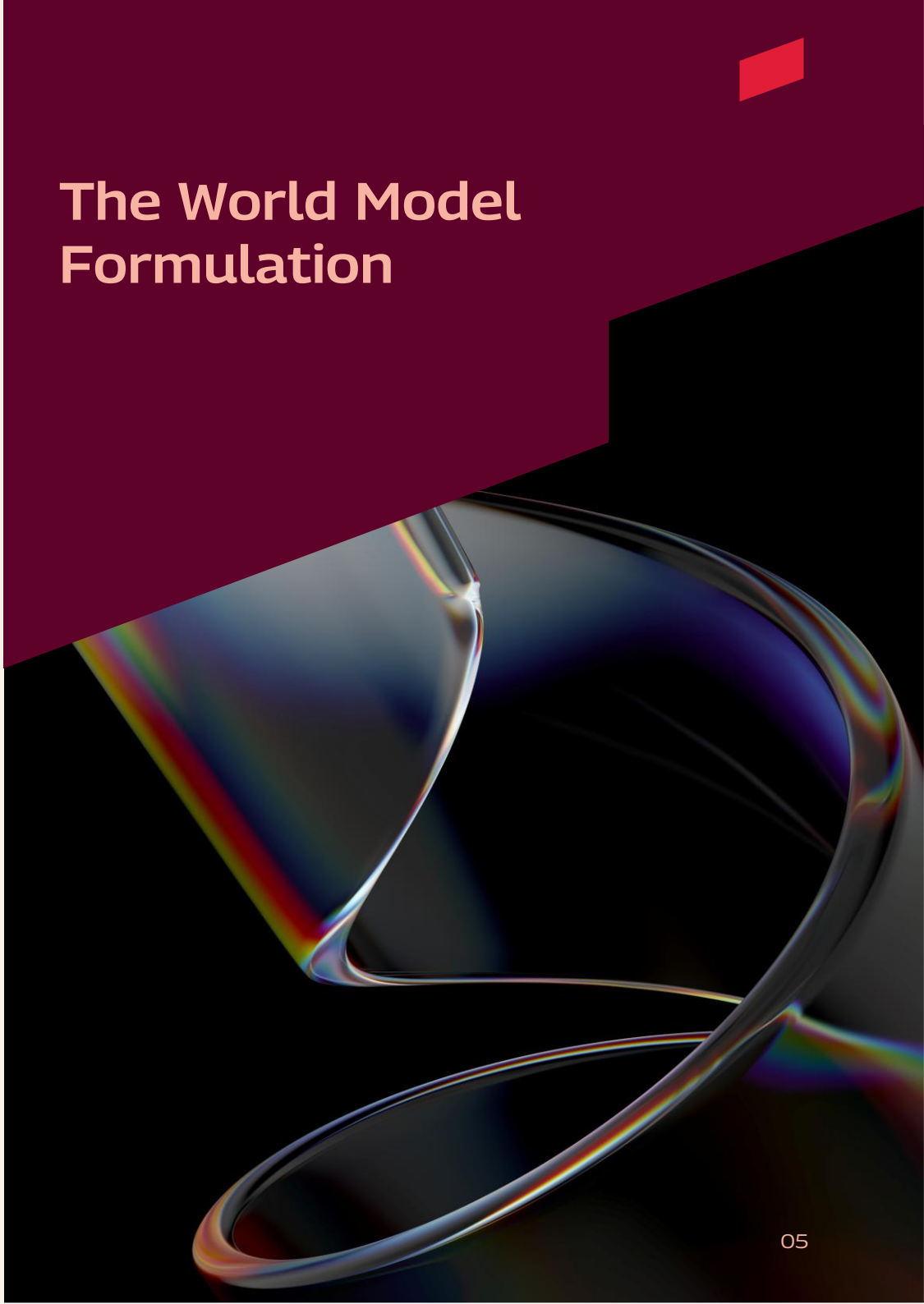
Today's AI excels at pretending to understand the world, but it doesn't. It predicts tokens, which do not reflect the nuances of reality.

The World Model Formulation

The combination of PreBEK and PoBEK helps humans to construct a coherent world model. This model serves as an internal mental map of how the world works, including its objects, cause-and-effect relationships, and social rules. It acts as a central reference system that guides every action human beings take.

This framework bears a striking resemblance to concepts in deep reinforcement learning (RL). In this analogy, a human can be viewed as an intelligent agent operating in the world. Every human lives their life through a set of actions in pursuit of a desired outcome, and when these outcomes are achieved, each individual is rewarded with dopamine, which determines satisfaction. Just like an agent embedded with a defined action and reward mechanism.

Moreover, the human world model is not static.
It continuously evolves through ongoing interaction with the environment, integrating a rich stream of sensory data from the five senses. Every action taken in response to this incoming data produces observable effects in the world, which are then fed back as new observations. This closed-loop interaction allows the internal model to be constantly updated, refined, and expanded throughout life.



The formulation of this world model is inherently stochastic. No two individuals share identical models. Each of the roughly 8 billion humans on Earth possesses a unique world model shaped by their personal combination of PreBEK and PoBEK and continuously updated through real-world interactions. This individualized, experience-driven evolution explains both the diversity of human perspectives and the remarkable adaptability of human intelligence.

Learning in humans is fundamentally prediction-error driven. It rarely occurs through the linear accumulation of knowledge. Instead, it is triggered when reality deviates from the expectations of the internal world model. When the world behaves exactly as predicted, there is little new learning and minimal dopamine. True learning emerges from surprise, the mismatch between the brain's predicted next state and the actual outcome.

To illustrate this, let's consider a game of cricket with a batsman. The individual's internal world model predicts the next delivery from the bowler. The bowler bowls, but the delivery is not what was expected. Such a rapid change in the internal model triggers the art of maneuvering and adapting to the given situation. This, in turn, strengthens human intelligence.

> ***Human intelligence is built on surprise, adaptation, and causal understanding. AI must learn the same way.***

The Opportunity: How Can AI Learn

To replicate the human learning framework for building a truly dynamic AI system, we must move beyond static neural architectures and adopt an internal world-model-based system. Currently, the AI systems lack sensory experience. They perceive the world in numerals. If systems are deliberately built by embedding symbolic structures and dynamic adaptation mechanisms within deep reinforcement learning frameworks, we can create effective AI systems.

Dream.AI, a concept in which AI learns by dreaming and predicting the next state of the world, can be extended to support this pursuit. The foundational idea behind Dream.AI, introduced in 2024, was to enable AI to learn a world model through observation and then simulate future states via an internal 'dreaming' or counterfactual reasoning process.2 This allowed the agent to perform wishful thinking and explore possible next states before acting in the real world. While innovative, Dream.AI still treated the learned world model as largely static. With a crucial advancement to this framework, especially with dynamicity, the system can continuously update its world model as the environment changes, enabling faster adaptation. This means the transition dynamics P(St+1|St), along with the action taken, are no longer static but evolve in real time.

Limitations of Current Deep RL and the Imperative for Dynamic Models

Current deep RL models remain largely static. Once trained in a fixed environment, they are unable to adapt quickly to changing conditions. Yet the real world is inherently dynamic: it continuously generates new patterns, rules, and challenges that demand rapid detection, reaction, and behavioral adjustments. This adaptive process is known as world model induction, where it is imperative that the AI agent learns as it interacts with the world. More importantly, there is a critical preparatory step that precedes full induction: the agent must first possess a flexible, symbolically grounded representation of the world that can be rapidly modified in response to change.

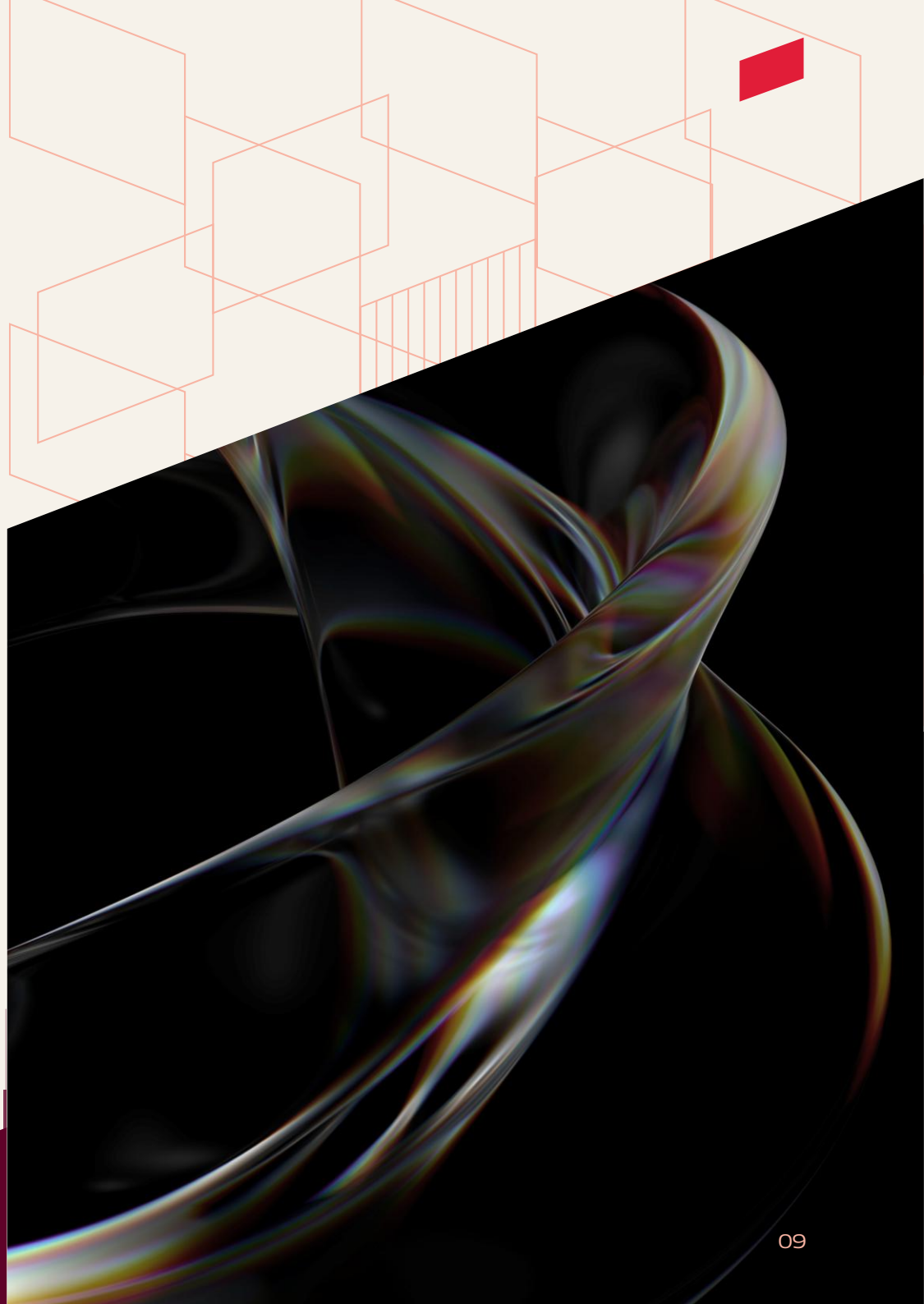
As deep RL with explicit, dynamic model representation becomes the need of the hour, it is essential to develop an architecture that reflects the real world. Although the physical world is three-dimensional, we can implement 2D Cartesian space for clarity and tractability. To test this in practice, we can consider video games. They provide clear state transitions and observable rewards, allowing us to evaluate the RL system's capabilities.

Advent of Symbolic Deep Reinforcement Learning

To prove this concept of model enhancement, we can build on the limitations of static RL models. As a primary target, we can consider 2D games, especially the ARC-AGI-3 challenge introduced by François Chollet and the ARC Prize team.3 ARC-AGI-3 is an interactive benchmark consisting of hundreds of novel, handcrafted game environments. In these environments, agents receive neither the rules nor the objective of each game. They must explore, infer goals on the fly, construct an internal model of the environment's dynamics, and discover effective strategies through interaction alone. The adoption of this foundation is critical to symbolic deep RL.

Why Symbolic Reasoning for Reinforcement Learning?

Symbolic representations lie at the heart of human cognition. Our PreBEK and PoBEK are largely symbolic in nature. It keeps us grounded, and based on this symbolic reasoning, we interact with the world, receive rewards, and live our lives. Symbolic deep RL, therefore, combines deep reinforcement learning with innate symbolic priors and dynamic world-model adaptation. This hybrid approach creates an AI capable of rapid, human-like adjustment to changing environments.



The Markov Decision Process (MDP) Framework and the Role of Symbolism

In the model-free architecture of deep RL, an environment is typically formalized as an MDPP defined by the tuple ⟨S, A, P, R, γ⟩, where:

Symbol	Meaning	What it Represents
S	State	All possible situations the AI model can be in.
P	Transition Probability	P(s′ | s, a) The likelihood of moving from one state to another.
A	Action	The set of available actions the AI model can take.
R	Reward	(Rss') Feedback received after taking an action.
Y	Discount	Balances immediate versus future rewards. (0 ≤ γ < 1)
Gt	Expected Return	The total future reward starting from timestep (t)

Formally, the cumulative return here is defined as:
Gt=Rt+1+ Ỵ(Rt+2)+ Ỵ^2 (Rt+3)+ Ỵ^3 (Rt+3).....
With immediate reward, it becomes **Rt+1.** Therefore, the state-value function is given as **v(s) = E(Gt | St=s, At=a)**

The Reinvention of Reward

In open-ended settings, the traditional scalar reward can no longer be trusted or directly applied. Since the objective is unknown and conditions are unclear, the reward signal must be verified and validated from its principles. This is where symbolism becomes essential in SDRL. Rather than depending on externally provided scalar values, such as game scores or points, the reward is now derived from observable changes in the environment state through a mechanism called the symbolic state differentiation.

Additionally, the reward is redefined based on what changed from the previous state, specifically, whether the screen vectors changed after a specific action or not. Instead of relying on predefined scores, the reward function depends on the degree of environmental change. The agent asks: did the screen change? Did any attempts or indicators reduce? This shift from scalar values to state differentiation allows the agent to learn without knowing the game's objective in advance.

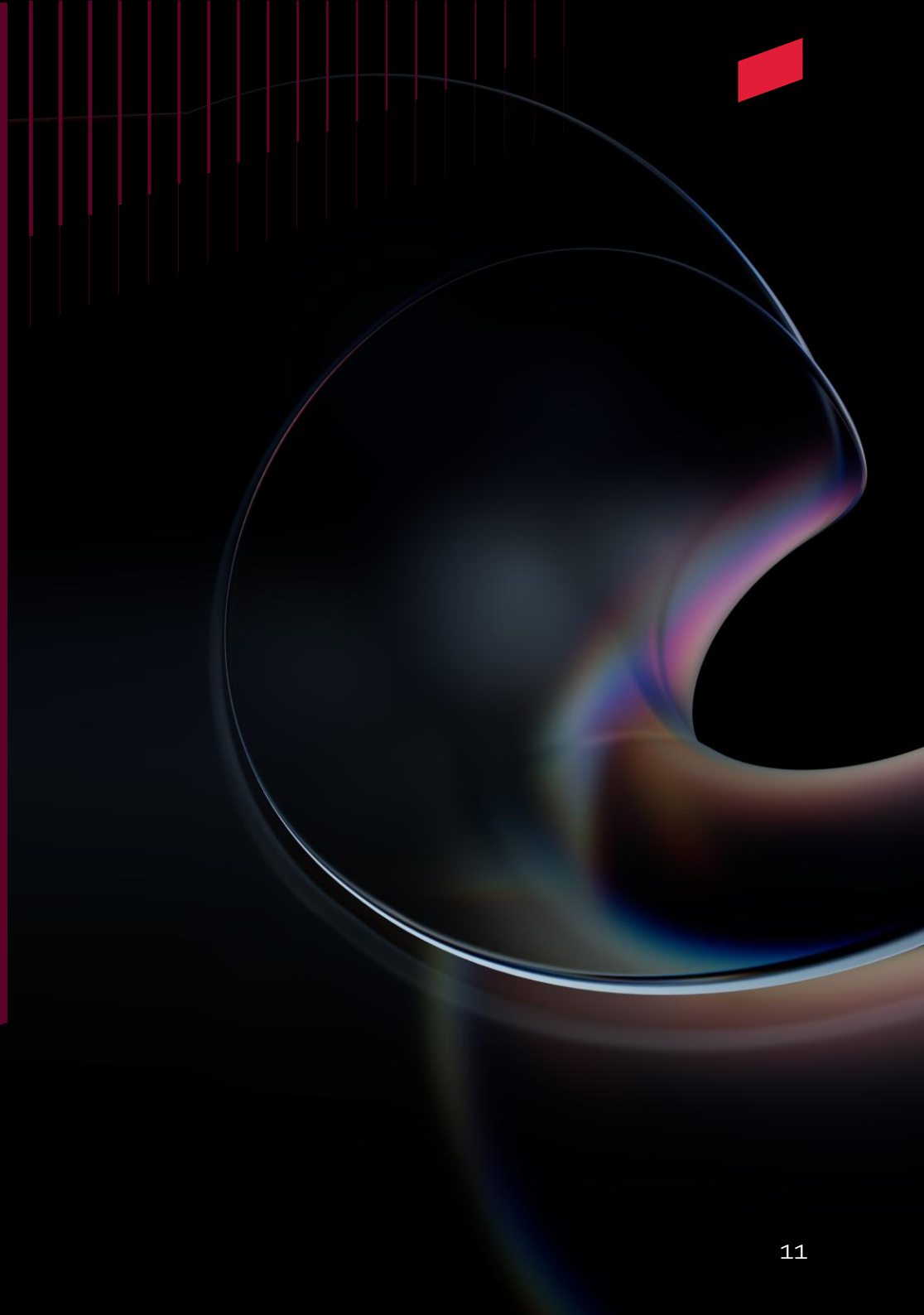
Symbolic Priors:
Embedding PreBEK in SDRL

A natural question arises: where exactly does the symbolism reside in Symbolic Deep RL?

To put it succinctly, it is explicitly embedded as priors before the agent begins learning, mirroring the role of PreBEK in human cognition. These symbolic priors are injected into the agent at initialization, providing it with foundational inductive biases about the structure of 2D environments.

In the context of 2D Cartesian games, the key PreBEK symbolic priors include:

- **Bounded Domain:** The agent is initialized with the understanding that it operates within a finite spatial region, typically a rectangle, circle, or grid. While the precise dimensions are unknown at the start, the agent uses a convolutional neural network (CNN) to quickly detect and internalize the boundaries from the initial observations. This prior establishes the playable area and prevents exploration of invalid spatial assumptions.



- **Actions:** The agent begins with compact, symbolically defined actions in the 2D plane, categorized as:

  <Action1>: Left

  <Action2>: Right

  <Action3>: Up

  <Action4>: Down

  <Action5>: Diagonal R-up, which is achieved by one block right up

  <Action6>: Diagonal L-up, which is achieved by one block left up

  <Action7>: Diagonal R-down, which is achieved by one block right and then down

  <Action8>: Diagonal L-down, which is achieved by one block left and then down

  <Action9>: Mouse click

  <Action10>: Spacebar

  <Action n>: Any key press

These actions are either prebuilt into the game, as we discover from the action space, or learned; the agent starts with pre-embedded knowledge of the actions it must execute to clear the game.

- **Bounce Boundary:** The agent cannot be initiated unless they are constrained by a bounce boundary, which is a bounded domain. The agent must learn to interact with these boundaries based on the state.
- **Reward:** The reward is the most critical component in SDRL. It is computed directly from the state change caused by the agent's action. Specifically, after each action, a reward signal is generated based on the vector differentiation between the current state and the next state of the environment.

Finally, vector differentiation quantifies the change between the current and next states of the environment. Large changes result in high positive or negative rewards, depending on the direction and context of the change.
The magnitude of this change serves as a proxy for progress: small changes indicate the agent is navigating through the environment, while significant changes suggest major advancement or task completion. The final reward signal combines both a scalar value and the vector of state change.

R = P.Sd (state vector differentiation). This is the change in value between the two states.

Now the state-value function becomes the expectation of reward:

$V\pi(s)=E\pi[(Rt+1).Sd+ \gamma V\pi(St+1)|St=s]$

The action value function subsequently becomes

$q\pi(s,a)=E\pi[(Rt+1).Sd+ \gamma q\pi(St+1,At+1)|St=s,At=a]$

Illustrative Example

Consider the following game environment, where neither the objective nor any scoring mechanism is provided to the agent.

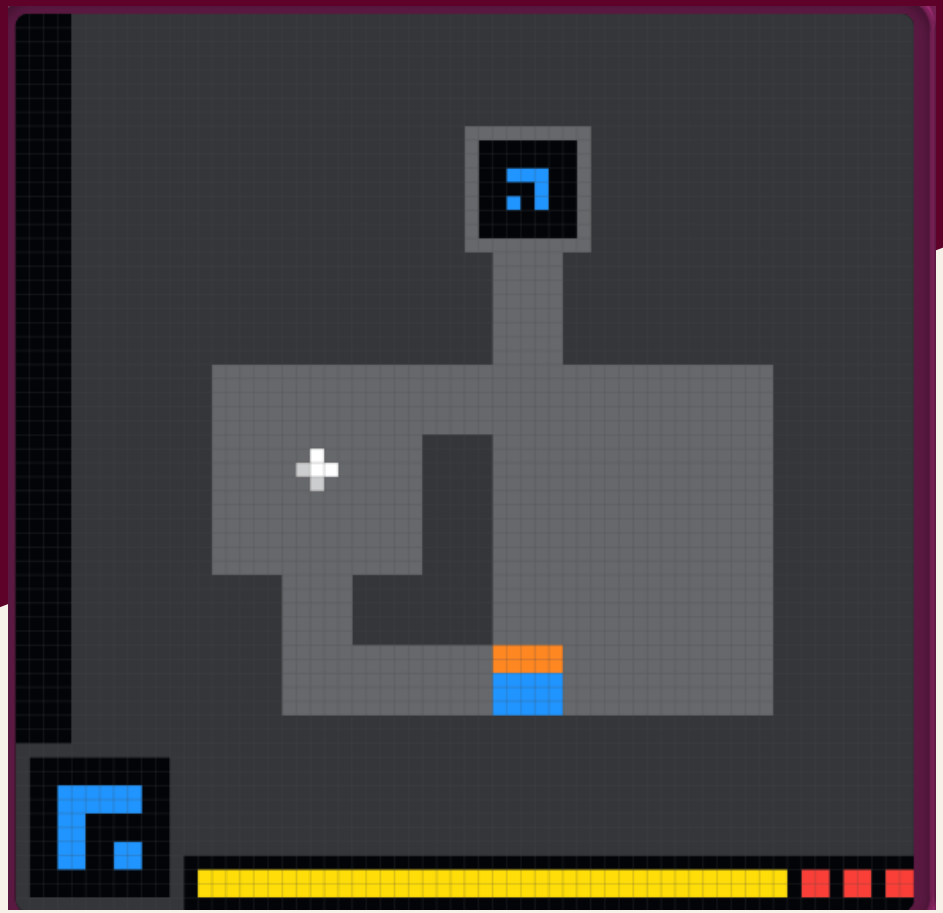


Figure 01: Illustrative 2D Game

The only available actions allow the agent to move the orange-blue block up and down. There is no visible score, no instructions, and no explicit goal.

- Through interaction and observation, the underlying objective can be discovered
- The blue-orange block must be moved forward
- There is a white cross that needs to be collected or eaten
- Once the cross is collected, the block must reach the black box at the top and sit inside it

This example perfectly demonstrates why symbolic state differentiation is powerful. Even without any external reward signal or objective description, meaningful progress is observable through changes in the environment. These symbolic events emerge naturally through vector differentiation between consecutive frames. SDRL leverages exactly this kind of intrinsic, observation-driven feedback to allow the agent to infer goals and learn effective policies with minimal data.

The Final Word

SDRL represents a fundamental shift from today's dominant paradigms of statistical pattern matching and static reinforcement learning. By embedding PreBEK-inspired symbolic priors, enabling dynamic world-model adaptation, and introducing symbolic state differentiation as an intrinsic reward mechanism, SDRL addresses the core weaknesses that have long plagued both LLMs and conventional deep RL systems.

SDRL does not merely simulate intelligence; it seeks to replicate the mechanisms that make human cognition so remarkably efficient. The framework is specifically designed to excel in challenging environments, where agents must discover rules, objectives, and strategies with zero prior information.

The implications are profound. Future AI systems built on SDRL principles will require far fewer training examples, adapt in real time to novel situations, and exhibit genuine causal understanding of the worlds in which they operate.

The era of static models is ending. The next generation of AI must be dynamic, symbolic, and continually learning, and that's exactly what SDRL delivers.

End Notes

1. Nawaz, U., Anees-ur-Rahaman, M., & Saeed, Z. (2025). A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. Intelligent Systems with Applications, 26, 200541. https://www.sciencedirect.com/science/article/pii/S2667305325000675?via%3Dihub

2. Analytics India Magazine. (n.d.). What is Nikhil Malhotra's dream AI? https://analyticsindiamag.com/ai-features/what-is-nikhil-malhotras-dream-ai

3. ARC Prize. (2026). ARC-AGI-3. https://arcprize.org/arc-agi/3

Nikhil
Malhotra

Chief Innovation Officer
& Global Head - AI,
Tech Mahindra

About the Author

Nikhil has been a researcher all his life and is now leading the growth of AI and Quantum Computing research within Tech Mahindra. His area of business research is how quantum Computing, AI, and neuroscience would inspire the growth of AI and the next change in society, business, and humanity. He has won numerous awards, including the 2020, 2021, and 2023 Innovation Congress awards, for being the most innovative leader in India.

Nikhil is also a TEDx speaker and the author of a best-seller book - Courage, the Journey of an Innovator. One of his long-standing visions has been to enable machines to talk in the local Indian dialects. Most notably, he has spearheaded Project Indus, Tech Mahindra's seminal effort to build Indic LLM (homegrown large language model), which was successfully launched globally in June 2024.Nikhil holds a master's degree in computing with a specialization in distributed computing from the Royal Melbourne Institute of Technology, Melbourne, and is an avid physicist.

About **Tech Mahindra**

Tech Mahindra (NSE: TECHM) offers technology consulting and digital solutions to global enterprises across industries, enabling transformative scale at unparalleled speed. With 147,000+ professionals across 90+ countries helping 1100+ clients, Tech Mahindra provides a full spectrum of services including consulting, information technology, enterprise applications, business process services, engineering services, network services, customer experience & design, AI & analytics, and cloud & infrastructure services. It is the first Indian company in the world to have been awarded the Sustainable Markets Initiative's Terra Carta Seal, which recognizes global companies that are actively leading the charge to create a climate and nature-positive future. Tech Mahindra is part of the Mahindra Group, founded in 1945, one of the largest and most admired multinational federation of companies. For more information on how TechM can partner with you to meet your Scale at Speed™ imperatives, please visit https://www.techmahindra.com/.

*Figures as per Q4, FY 26.

www.techmahindra.com
www.linkedin.com/company/tech-mahindra
www.x.com/Tech_Mahindra