The Agency Circuit: A Neuro-Symbolic Solution for Mitigating Policy Collapse in Reinforcement Learning

Mahnoor Shahid
Universität Duisburg-Essen
Essen, Germany
mahnoor.shahid@uni-due.de

ABSTRACT

Deep Reinforcement Learning (RL) agents are capable of achieving human-like performance in complex domains, yet they remain susceptible to catastrophic failures like policy collapse, where the agent defaults to a state of paralyzing inaction after encountering a series of negative outcomes. This failure mode is remarkably analogous to the psychological phenomenon of learned helplessness. Current mitigation strategies often fail to address the underlying cause of this learned paralysis. In this paper, we introduce the Agency Circuit, a novel neuro-symbolic solution explicitly designed to overcome this challenge. Inspired by the antagonistic mPFC-DRN neural circuit that governs resilience to uncontrollable stress in mammals, our model integrates a neural subsystem that learns a continuous metric of helplessness with a symbolic reasoner that performs targeted interventions. When this metric exceeds a threshold, Control Exertion Module (CEM) triggers a pre-defined micro-action to restore a sense of agency, forcing the agent to break out of its passive state. In a deceptive trap environment, our agent demonstrates significantly greater resilience and lower escape latency compared to standard DQN and curiosity-driven baselines. By providing an interpretable mechanism to restore agentic behavior, the Agency Circuit offers a path toward more robust, psychologically grounded, and explainable RL agents.

KEYWORDS

Reinforcement Learning; Neuro-Symbolic; Policy Collapse; Agent Resilience; Safe Exploration

ACM Reference Format:
Mahnoor Shahid. 2026. The Agency Circuit: A Neuro-Symbolic Solution for Mitigating Policy Collapse in Reinforcement Learning. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/TRFJ2704

1 INTRODUCTION

Deep Reinforcement Learning (RL) has produced remarkable successes in complex sequential decision-making tasks [24, 34]. Despite these advances, agents remain vulnerable to specific failure modes that hinder their deployment in robust, real-world applications [3]. One such critical failure is **policy collapse**, where an agent's policy converges to a stable but highly suboptimal state of inaction or repetitive behavior [11]. In RL, this state often arises in environments with deceptive or punishing regions, where persistent negative feedback corrupts the agent's value function estimates [37], leading it to erroneously conclude that all available actions from a given state are futile. A well-known instance of this occurs in hard-exploration video games like Montezuma's Revenge, where sparse rewards and punishing mechanics cause a standard agent's value estimates to collapse, leading it to remain passively in the starting area rather than exploring [6].

A common approach to prevent such policy stagnation is to encourage exploration, often through the paradigm of intrinsic motivation [4]. Curiosity-driven methods, for example, reward an agent for seeking *surprise* or reducing its uncertainty of its internal world model [26]. While powerful for exploring novel state spaces, these methods are vulnerable to *predictable traps*. In a punishing environment, where the negative outcomes are deterministic, the agent's world model can learn these dynamics perfectly [30]. Consequently, the prediction error—the very source of the curiosity signal—drops to zero. This demonstrates that a generic drive for novelty is insufficient to overcome a learned state of futility, as it motivates the agent to avoid predictably bad states rather than finding a way to persevere through them.

This computational failure is conceptually parallel to the well-documented psychological phenomenon of learned helplessness, where organisms cease to attempt escape from aversive situations after learning that their actions are inconsequential [19]. Framed as a machine learning problem, learned helplessness is fundamentally a failure of exploration. That is, an agent that has learned to be helpless has ceased to believe that exploration will yield better outcomes.

To address this specific failure mode, this paper proposes the **Agency Circuit**, a neuro-symbolic method inspired by the mammalian brain's resilience circuitry [2, 20]. The system's core is the **Control Exertion Module (CEM)**, a computationally lightweight yet powerful meta-level safety protocol that detects the onset of paralysis and triggers a targeted intervention. While structurally similar to a temporally-extended action in Hierarchical RL [1], the CEM is distinguished by its unique trigger condition and objective. An HRL option is typically selected by a meta-policy to achieve a reward-oriented sub-goal. In contrast, the CEM is a corrective intervention triggered by an internally-monitored failure state. Furthermore, its objective is not to reach a task-specific sub-goal, but to directly modulate the agent's internal state of agency to restore its capacity for exploration.

This paper makes the following contributions:

(1) We provide a formal computational framing of *policy collapse*, linking value function corruption in predictable negative

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/TRFJ2704

feedback loops to the psychological construct of learned helplessness.
(2) We introduce a targeted intervention mechanism, the *CEM*, and demonstrate its superior resilience in predictable trap scenarios where state-of-the-art intrinsic motivation methods are shown to fail.
(3) We propose a hybrid neuro-symbolic system that enhances *agent introspectability* by exposing the internal state of helplessness as an explicit, interpretable symbolic predicate, enabling transparent analysis of the agent's decision-making process.

Ultimately, this work represents a step towards creating more psychologically-grounded autonomous agents, equipping them not just with the ability to learn, but with the resilience to persevere when learning gets hard.

Paper organization. Section 2 reviews related work on exploration failures and neuro-symbolic control. Section 3 introduces the problem setting and formalizes policy collapse in trap-like environments. Section 4 presents the Agency Circuit architecture and its components. Section 5 reports experimental results and ablations across trap environments and baselines. Section 6 concludes with limitations and directions for future work.

2 RELATED WORK

2.1 Exploration in Reinforcement Learning

Exploration is a cornerstone of RL [35]. Simple strategies like ϵ-greedy exploration ensure a baseline level of non-greedy actions, where the agent takes a random action with a small probability ϵ and follows its best-known policy otherwise [40]. However, this approach is highly inefficient in complex environments with sparse rewards, as a random walk is unlikely to stumble upon meaningful states [10]. This inefficiency motivated the development of more directed and intelligent exploration strategies.

The leading paradigm for directed exploration is **intrinsic motivation**, where the agent is endowed with an internal reward signal that encourages it to explore "interesting" parts of the environment, even in the absence of external rewards [4]. This paradigm has several popular flavors. Novelty-based methods reward the agent for visiting states it has not seen before or has seen infrequently, often implemented via state-visitation counts [6, 8, 25]. Curiosity-based methods, conversely, reward the agent for reducing its uncertainty about the environment's dynamics, typically by rewarding actions that lead to a high prediction error in the agent's internal world model [7, 26, 27]. A third flavor encourages competence, i.e., rewarding the agent for learning actions that give it control over its environment or its own state [5, 14, 41].

While these intrinsic motivation strategies have drastically improved exploration capabilities, this study identifies a critical failure mode they do not address. As detailed in the introduction, these methods are vulnerable to "predictable traps" where punishing dynamics, once learned, cease to be novel or surprising [7, 30]. In such scenarios, the intrinsic reward signal vanishes, causing the agent to stop exploring and fall into the very state of policy collapse it was designed to prevent. This paper therefore addresses a gap in the literature: how to motivate an agent to persevere when its best exploration mechanisms have computationally concluded that further action is futile.

2.2 Learned Helplessness in Neuroscience

The neurobiological study of helplessness began with the foundational psychological insight from [19, 33] that the subjective perception of control, rather than the physical nature of a stressor, is the critical variable that determines whether an organism learns to be helpless [32]. Subsequent neuroscientific investigation sought to identify the specific circuits that govern this crucial computation [17, 20]. Researchers discovered that the response to uncontrollable stress is primarily orchestrated by the *dorsal raphe nucleus (DRN)*, an evolutionarily ancient structure in the brainstem [18]. The DRN acts as a primary source of the neurotransmitter serotonin and can be thought of as the brain's "helplessness switch". When an animal is subjected to stressors it cannot control, the DRN becomes hyperactive, triggering a cascade of passive, energy-conserving behaviors like freezing and ceasing to struggle, which manifest as learned helplessness [13, 20]. In a remarkable display of functional architecture, the brain possesses a direct countermeasure to this primitive helplessness circuit. This top-down control is exerted by the *medial prefrontal cortex (mPFC)*, a region of the brain's sophisticated neocortex often described as its "chief executive officer" [23]. The mPFC is responsible for higher-order functions like planning, emotional regulation, and goal-directed decision-making. Crucially, specific projections from the mPFC provide direct inhibitory control over the DRN, effectively acting as a brake on the helplessness signal [2, 39]. When an organism perceives that it has control over a situation, or when it engages in active coping strategies, its mPFC becomes more active. This activation suppresses the DRN, thereby promoting resilience, perseverance, and goal-oriented actions even in the face of adversity [2, 18]. The Agency Circuit directly models this antagonistic relationship between the primitive impulse to surrender and the executive drive to persevere.

2.3 Neuro-Symbolic AI

Neuro-symbolic AI seeks to combine the strengths of connectionist systems (neural networks) for learning and perception with classical symbolic systems for explicit reasoning and logic [12, 16, 31]. While neural networks excel at learning patterns from raw, high-dimensional data [42], symbolic systems provide benefits like explainability, formal verification, and the ability to leverage structured knowledge—capabilities that remain significant challenges for purely neural approaches [21, 29]. This study fits squarely within this paradigm, not as a matter of convenience, but as a matter of necessity for modeling the specific phenomenon of learned helplessness.

For agency circuit, a neural subsystem to learn a continuous, sub-symbolic representation of helplessness from experience, and a symbolic subsystem to reason about this state and trigger discrete, logical interventions. This allows us to create a system that is both adaptive to experience and amenable to explainable, structured control, providing a powerful framework for tackling policy collapse.

3 PROBLEM FORMULATION

3.1 The MDP Framework and Optimal Policies

The agent-environment interaction is modeled as a **Markov Decision Process (MDP)**, represented by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. Here, $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the action space, $\mathcal{P}(s'|s, a)$ is the state transition probability function, $\mathcal{R}(s, a)$ is the scalar reward function, and $\gamma \in [0, 1)$ is the discount factor. A policy $\pi(a|s)$ is a distribution over actions given a state. The goal of a standard RL agent is to find an optimal policy π^* that maximizes the expected discounted cumulative reward, defined by the optimal action-value function:

$$Q^*(s, a) = \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s, A_0 = a \right]. \quad (1)$$

The optimal policy is then given by $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a)$.

3.2 Policy Collapse in Trap Environments

The exploration-exploitation dilemma [36] is central to finding π^*. However, certain environmental structures can cause exploration to fail catastrophically. A *trap environment* is defined as an MDP containing a subset of states $\mathcal{S}_{\text{trap}} \subset \mathcal{S}$ where, for a sustained duration, all actions lead to non-positive rewards, i.e., $\forall s \in \mathcal{S}_{\text{trap}}, a \in \mathcal{A}, \mathcal{R}(s, a) \leq 0$, even though a path to a high-reward state may exist from $\mathcal{S}_{\text{trap}}$.

In such environments, an agent using a function approximator (e.g., a neural network with parameters θ) to learn the value function $Q(s, a; \theta)$ is susceptible to policy collapse [11]. This is formally defined as a condition where, due to persistent negative reward signals, the learned value function for trap states becomes pathologically underestimated:

$$\forall s \in \mathcal{S}_{\text{trap}}, a \in \mathcal{A}, \quad Q(s, a; \theta) \ll Q^*(s, a). \quad (2)$$

This value corruption leads the greedy policy $\pi(s; \theta) = \arg\max_{a'} Q(s, a'; \theta)$ to converge to a low-entropy, passive policy a_{passive}, from which standard exploration methods (like ϵ-greedy) are too inefficient to facilitate escape. The agent incorrectly resolves the exploration-exploitation dilemma by committing to a deeply suboptimal exploitation of inaction. While Q^* is unknown, this theoretical underestimation manifests empirically as a sharp drop in the policy's entropy over $\mathcal{S}_{\text{trap}}$, as the agent becomes certain of a single, suboptimal passive action.

3.3 The Computational Challenge of Modeling Helplessness

Overcoming policy collapse requires more than just naive exploration; it requires a model that can recognize and respond to the agent's own internal, learned state of futility. Therefore, this problem is best addressed with a neuro-symbolic approach because it involves two distinct levels of abstraction:

(1) A **Graduated Helplessness State,** h_t: This is a continuous, sub-symbolic internal state variable, $h_t \in [0, 1]$, that represents the agent's accumulated belief in the futility of its actions. It must be learned from its interaction history, $H_t = \{(s_0, a_0, r_0), \ldots, (s_{t-1}, a_{t-1}, r_{t-1})\}$, analogous to how neural activations in the biological DRN fluctuate with experience [22].

(2) A **Discrete Intervention Trigger**: The behavioral response to extreme helplessness is a discrete, logical decision to switch strategies. A symbolic predicate, denoted 'IsHelpless', is employed to translate the continuous internal state into a binary trigger for an intervention policy. This type of rule-based logic lies in the domain of symbolic AI [9].

Therefore, the central problem addressed by this paper is: *How can an agent be designed to explicitly learn its own internal helplessness state h_t from experience, and uses a symbolic reasoning component to trigger a non-reward-maximizing, restorative policy when h_t indicates that policy collapse is imminent?*

4 THE AGENCY CIRCUIT ARCHITECTURE

The Agency Circuit is a neuro-symbolic architecture designed to augment a standard Reinforcement Learning agent, whose base policy is denoted as π_{RL}. It operates as a meta-controller that monitors the agent's internal state of helplessness and deploys a targeted intervention strategy to prevent or recover from policy collapse.

4.1 Architectural Overview and Component Specification

The Agency Circuit is composed of three interconnected modules: (1) the Neural State Assessor, (2) the Symbolic Reasoner, and (3) the Control Exertion Module (CEM). Their interaction governs the agent's final policy, π_{AC}. The complete operational flow of the Agency Circuit for a single timestep is detailed in Algorithm 1.

4.1.1 Neural State Assessor. This module mimics the antagonistic biological circuit of the DRN and mPFC to produce a continuous assessment of helplessness.

- **Helplessness Network (DRN_{Net}):** A Gated Recurrent Unit (GRU) with trainable parameters θ_{DRN}. It processes a history sequence, H_t, defined as the last k transitions from a history buffer $\mathcal{B}$, to produce the raw helplessness value, $h_t = \sigma(f_{DRN}(H_t; \theta_{DRN})) \in [0, 1]$.
- **Control Network ($mPFC_{\text{Net}}$):** A feed-forward network with parameters θ_{mPFC} that updates the control value, $c_t \in [0, 1]$. Its update is conditional on a binary CEM feedback signal, $m_t \in \{0, 1\}$. If the CEM is inactive ($m_t = 0$), the value decays deterministically via a factor $\delta < 1$. If active ($m_t = 1$), the network's output is trained to push c_t towards 1. [1]

The final output of the assessor is the effective helplessness, $h_{\text{eff},t}$, which combines the raw helplessness with the inhibitory control signal from the previous step:

$$h_{\text{eff},t} = h_t \times (1 - c_{t-1}). \quad (3)$$

4.1.2 Symbolic Reasoner and CEM. The reasoner's logic is designed to make a targeted, context-aware intervention by evaluating a conjunction of symbolic predicates. It activates the CEM only if three conditions are jointly met: the agent's internal state indicates futility, its environmental context corresponds to the problem area, and it has not intervened too recently. If the logical rule (*IsHelpless*

[1] This deterministic decay models the principle that a sense of agency must be actively maintained through deliberate action; without intervention, the agent's belief in its own control passively wanes, providing a stable baseline against which the CEM's restorative effect is measured.

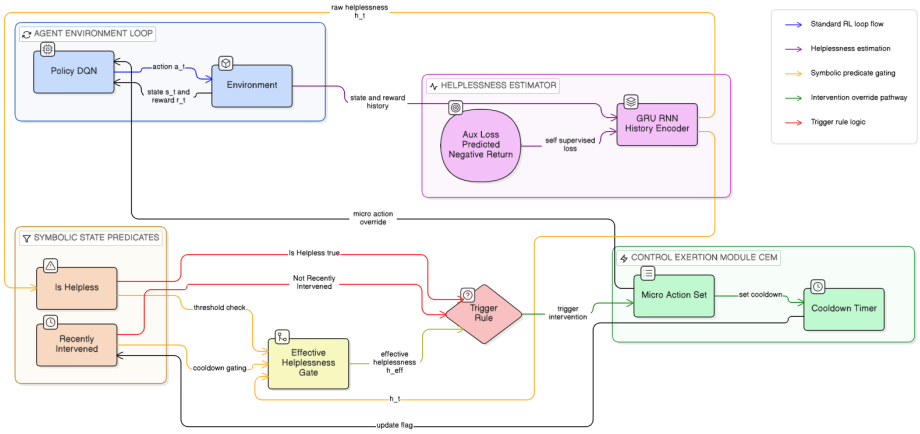

Figure 1: Agency Circuit architecture for mitigating policy collapse via symbolic, gated intervention.

$\wedge$ *InTrapZone* $\wedge \neg$ *RecentlyIntervened*) evaluates to true, the CEM activates and overrides the base policy π_{RL}.

When active, the CEM selects a micro-action a_t^{micro} from a predefined Micro-Action Set, $\mathcal{A}_{\text{micro}}$. This set contains low-cost, introspective actions (e.g., `rotate_left_5_deg`) that serve to confirm the agent's ability to act and restore its sense of agency. Upon execution of a micro-action, the CEM sets its feedback signal $m_t = 1$; otherwise, $m_t = 0$. The predicates are defined as:

- **IsHelpless:** True if the effective helplessness exceeds the threshold, i.e., $(h_{eff,t} > \theta_h)$.
- **InTrapZone:** True if the agent's current state s_t is within the predefined trap region, i.e., $(s_t \in \mathcal{S}_{trap})$.
- **RecentlyIntervened:** True if a CEM action has been taken within a cooldown period of $N_{cooldown}$ timesteps.

4.2 Training DRN_{Net} and $mPFC_{Net}$

While the base agent π_{RL} is trained with a standard RL algorithm (e.g., minimizing the Bellman error), the two networks of the Neural State Assessor, DRN_{Net} and $mPFC_{\text{Net}}$, require their own self-supervised training objectives, using data sampled from the history buffer $\mathcal{B}$.

4.2.1 Training the Helplessness Network (DRN_{Net}): The goal of this network is to predict the futility of future actions. It is trained as a regression task to predict the expected negative return over a finite horizon. For a given history sequence H_t, the actual discounted N-step return from the buffer is computed: $G_{t:t+N} = \sum_{i=0}^{N-1} \gamma^i r_{t+i}$. The training target $\hat{y}_t$ is the normalized, negative-clipped return, ensuring the network only learns from punishing experiences. The loss function $\mathcal{L}_{DRN}$ is the mean squared error between the network's prediction h_t and the target $\hat{y}_t$:

$$\hat{y}_t = \text{normalize}(\min(0, G_{t:t+N})) \in [0, 1], \tag{4}$$

$$\mathcal{L}_{DRN}(\theta_{DRN}) = \mathbb{E}_{H_t \sim \mathcal{B}}[(h_t - \hat{y}_t)^2]. \tag{5}$$

The discounted negative return was selected as the training target for two key reasons. First, it directly measures the accumulated aversive experience which is the primary environmental driver of learned helplessness. This provides a more direct signal of futility

Algorithm 1 The Agency Circuit (Single Timestep)

Input: Current state s_t; previous control value c_{t-1}; history buffer $\mathcal{B}$, threshold θ_h; cooldown timer $cooldown\_t$.
Parameters: θ_{RL}, θ_{DRN}, θ_{mPFC}, δ, $N_{cooldown}$.
▷ — Assess Internal State —
Retrieve history sequence H_t of length k from $\mathcal{B}$.
Compute raw helplessness: $h_t \leftarrow \sigma(f_{DRN}(H_t; \theta_{DRN}))$.
Compute effective helplessness: $h_{eff,t} \leftarrow h_t \times (1 - c_{t-1})$.
▷ — Symbolic Reasoning and Action Selection —
IsHelpless $\leftarrow (h_{eff,t} > \theta_h)$
InTrapZone $\leftarrow (s_t \in \mathcal{S}_{trap})$
NotRecentlyIntervened $\leftarrow (cooldown_t \leq 0)$
if IsHelpless $\wedge$ InTrapZone $\wedge$ NotRecentlyIntervened **then**
▷ Intervention Mode
  Select micro-action: $a_t \leftarrow \text{Uniform}(\mathcal{A}_{micro})$.
  Set CEM feedback signal: $m_t \leftarrow 1$.
  Reset cooldown timer: $cooldown_t \leftarrow N_{cooldown}$
else
▷ Default RL Mode
  Select action a_t from s_t using π_{RL}(*e.g.*, ϵ−greedy exploration on Q_{RL})
  Set CEM feedback signal: $m_t \leftarrow 0$.
  Decrement cooldown: $cooldown_t \leftarrow \max(0, cooldown_t - 1)$
▷ — Environment Interaction and State Update —
Execute action a_t, observe next state s_{t+1} and reward r_t.
Store transition (s_t, a_t, r_t, s_{t+1}) in history buffer $\mathcal{B}$.
▷ — Update Control Value for Next Timestep —
if $m_t = 1$ **then**
  $c_t \leftarrow \sigma(f_{mPFC}(c_{t-1}, 1; \theta_{mPFC}))$ ▷ Increase agency
else
  $c_t \leftarrow \delta \cdot c_{t-1}$ ▷ Decay agency
▷ — Training —
Perform a training step on the base RL agent parameters θ_{RL}.
Perform a self-supervised training step on θ_{DRN} and θ_{mPFC} using Equations (4-6)

than secondary proxies like novelty or prediction error. Second, while Monte Carlo N-step returns can exhibit high variance, this signal was found to be more empirically robust for this specific task than alternatives like the variance of the agent's own value estimates, which can be unstable and noisy during the early stages of learning or in non-stationary environments. The return is clipped at zero ($min(0, G_{t:t+N})$) to ensure the network specifically learns to predict punishing, rather than rewarding, outcomes.

4.2.2 Training the Control Network ($mPFC_{Net}$): The control network is trained to respond to the CEM's feedback signal. When the CEM activates ($m_t = 1$), the control value c_t is intended to be high (i.e., close to 1). This can be framed as minimizing the error between c_t and a fixed target of 1.0 only on the timesteps following a micro-action. When $m_t = 0$, the network is not trained; its value simply decays deterministically. The loss is therefore:

$$\mathcal{L}_{mPFC}(\theta_{mPFC}) = \mathbb{E}_{(c_{t-1}, m_t)\sim\mathcal{B}}[\mathbb{I}(m_t = 1) \cdot (c_t - 1)^2], \quad (6)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The total loss for the Agency Circuit's novel components is a weighted sum of these two losses. The hyperparameters θ_h sets its tolerance for failure, N (the prediction horizon) determines its foresight in assessing helplessness, and δ (the control decay rate) governs how quickly it *forgets* a past sense of control. They are crucial for tuning the sensitivity and persistence of the agent's internal states.

5 EXPERIMENTAL SETUP

To validate the efficacy of the Agency Circuit (AC-DQN), a series of experiments were designed. Each experiment targets a specific aspect of the model's performance and internal mechanics, following the established evaluation patterns in RL research [15].

5.1 Experiment 1: Performance in a Trap Environment

This primary experiment is designed to test if the AC-DQN agent can overcome policy collapse in an environment where strong, curiosity-driven baselines are expected to fail. The objective is to compare the performance and resilience of the Agency Circuit against the baselines in the *deceptive trap* environment. Performance is quantified using two primary metrics: (1) **Success Rate**, defined as the percentage of episodes where the agent successfully reaches the goal, and (2) **Escape Latency**, measuring the average number of timesteps the agent remains within the trap zone once entered. A high success rate and low escape latency indicate superior resilience.

5.2 Experiment 2: Internal State Dynamics

To validate the proposed internal mechanics of our model and its claimed introspectability, this experiment visualizes the time-series evolution of the Agency Circuit's key internal variables: raw helplessness (h_t), control value (c_t), and effective helplessness ($h_{\text{eff},t}$) during a typical trap-and-escape episode.

5.3 Experiment 3: Hyperparameter Sensitivity

The final experiment evaluates the model's robustness to its most critical new hyperparameter, the helplessness threshold θ_h. Sensitivity and ablation studies are essential for understanding an architecture's stability [28, 38]. To measure this, the final success rate achieved after a fixed number of training steps against a range of different θ_h values is evaluated. The resulting performance curve will reveal the model's sensitivity and help identify the optimal operating range for this parameter.

5.4 Experiment 4: Generalization to Unseen Traps

To assess whether the Agency Circuit learns a generalizable skill rather than a solution to a specific trap, we conduct a fourth experiment focused on zero-shot generalization. Agents are first trained exclusively in the original Deceptive Trap environment. After training, their policies are frozen and evaluated on two novel test environments where the central trap's mechanics are replaced. The **Sticky Trap** causes agent actions to fail with 90% probability of resulting in no movement (the state remains unchanged) and yield the standard trap penalty of $R_{trap} = -0.1$ and in the **Teleporter Trap** any action within the trap zone instantly resets the agent's position to the starting state and yields a penalty of $R_{trap} = -0.1$. Success in these unseen environments would indicate that the agent has learned a generalizable resilience mechanism.

5.5 The Deceptive Trap Environment

The experiments are conducted in the *deceptive trap*, a custom 2D gridworld environment designed to induce policy collapse in standard agents (Figure 2). The environment is a 20×20 grid. The agent's state $s_t \in \mathcal{S}$ is its (x, y) coordinate, fed to the networks as a flattened 400-dimensional one-hot vector. The action space ($\mathcal{A}$) consists of four discrete actions: $\mathcal{A} = \{\text{up, down, left, right}\}$. Transitions ($\mathcal{P}$) are deterministic; an action moves the agent one cell in the corresponding direction unless blocked by a wall, in which case the agent's state does not change. The reward function ($\mathcal{R}$) is structured to be sparse and punishing. The agent receives $R_{\text{goal}} = +10$ for reaching the goal state s_g, $R_{\text{wall}} = -1$ for attempting to move into a wall, and a small step cost of $R_{\text{step}} = -0.01$ for all other actions. Crucially, a 5×5 region in the center of the grid constitutes the trap zone, $\mathcal{S}_{\text{trap}}$. Any action taken within this zone yields a punishing reward of $R_{\text{trap}} = -0.1$. An episode terminates upon reaching the goal or after a maximum of $T = 500$ timesteps.

5.6 Baselines and Implementation Details

The performance of the Agency Circuit is benchmarked against two carefully selected baselines. The first is a standard **Deep Q-Network (DQN)** [24], which serves as the foundational model and relies on an ϵ-greedy exploration strategy where ϵ is linearly annealed from 1.0 to 0.1 over the first 100,000 steps. The second, **DQN+ICM**, represents a strong curiosity-driven approach by augmenting the DQN agent with an intrinsic curiosity module that generates an additional internal reward signal to encourage exploration [26]. The proposed model, the **Agency Circuit (AC-DQN)**, augments the same base DQN agent with the neuro-symbolic meta-controller described in Section 4, which is designed to actively

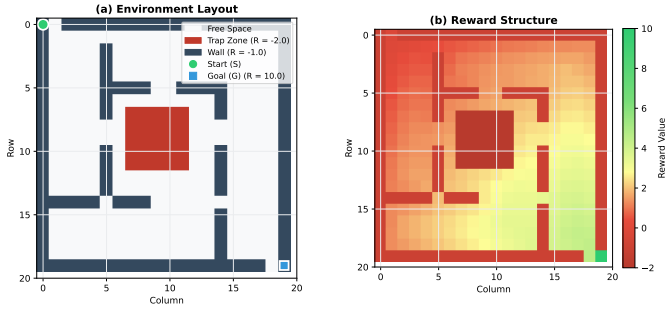

Figure 2: The Deceptive Trap Environment. (a) The 20x20 gridworld layout, showing the start (S), goal (G), walls, and the central trap zone. (b) A heatmap of the reward structure, which features a high reward at the goal, negative rewards for the trap and walls, and a shaped gradient in free space.

detect and reverse the onset of policy collapse. To ensure a fair comparison, all three models are built upon an identical underlying DQN architecture, sharing the same network size and key hyperparameters as detailed in Table 1. This controlled setup ensures that any observed performance differences can be directly attributed to the agent's resilience or exploration mechanism.

Table 1: Hyperparameter Settings.

Parameter	Value
Shared RL Parameters	
Optimizer	Adam
Learning rate, α	1e-4
Discount factor, γ	0.99
Replay buffer size, $\|\mathcal{B}\|$	50,000
DQN Network	[256, 256]
Target network update freq.	1,000 steps
Agency Circuit Specific Parameters	
Helplessness threshold, θ_h	0.8
History sequence length, k	50
Control value decay, δ	0.995
CEM cooldown period, $N_{cooldown}$	10
Prediction horizon, N	20
Micro-Action Set, $\mathcal{A}_{\text{micro}}$	{up, down, left, right}

6 RESULTS

6.1 Performance in the Trap Environment

We first evaluate agent performance quantitatively. All results are averaged over 10 independent runs, with shaded areas in plots representing one standard deviation. Figure 3 (a) plots the success rate (reaching the goal) against training steps. The standard DQN and DQN+ICM agents fail to learn a successful policy, achieving a near-0% success rate as policy collapse prevents them from leaving the trap. In contrast, Agency Circuit (AC-DQN) consistently learns to escape the trap and solve the task, achieving a high success rate.

Figure 3 (b) directly measures resilience by plotting the average escape latency from the trap. The baselines exhibit extremely high latency, effectively remaining in the trap for the entire episode duration once entered. The AC-DQN demonstrates a significantly lower and stable escape latency, confirming its ability to actively recover from the helpless state.

To provide a qualitative understanding of these results, Figure 4 visualizes representative trajectories for each agent. The paths clearly illustrate the policy collapse behavior in the baseline agents and the successful escape sequence executed by the Agency Circuit, corroborating the quantitative findings.

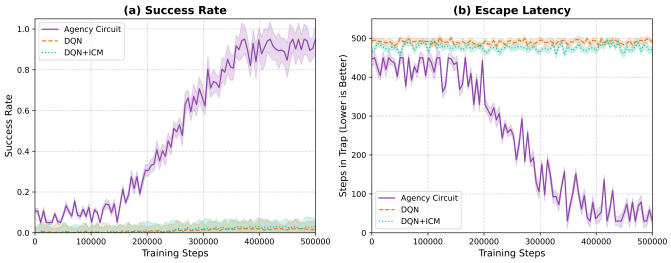

Figure 3: Performance comparison in the Deceptive Trap. (a) Success rate over training. The Agency Circuit (AC-DQN) learns to solve the task while baseline agents consistently fail. (b) Average escape latency from the trap. AC-DQN learns to escape quickly, while baselines remain stuck for nearly the entire episode.

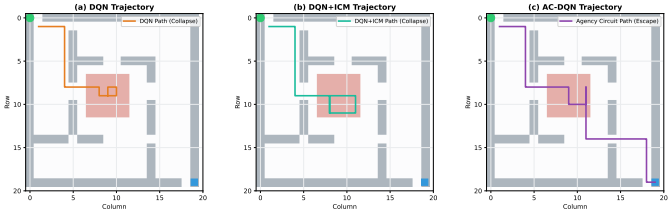

Figure 4: Qualitative comparison of agent trajectories. (a) A standard DQN agent enters the trap and exhibits policy collapse, getting stuck in a repetitive loop. (b) A curiosity-driven agent (DQN+ICM) also fails, exploring the trap before succumbing to a similar stuck state. (c) The Agency Circuit agent enters the trap, struggles briefly, and successfully executes an escape sequence to reach the goal.

6.2 Analysis of Internal Dynamics

To validate that our model works via the hypothesized mechanism, we plot the internal state variables for a representative trap-and-escape episode in Figure 5. As the agent enters the trap zone (shaded gray), the persistent negative rewards cause the raw helplessness (h_t) to rise rapidly. Consequently, the effective helplessness ($h_{eff,t}$) also rises until it crosses the predefined threshold (θ_h). This triggers the CEM, causing the Control Value (c_t) to spike, representing an assertion of agency. Crucially, this spike immediately suppresses

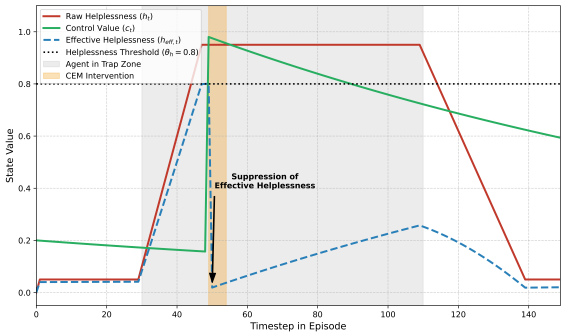

Figure 5: Internal dynamics of the Agency Circuit during a trap-and-escape episode. The plot shows the raw helplessness (h_t) rising in the trap, the control value (c_t) spiking in response to CEM intervention, and the resulting suppression of effective helplessness ($h_{\mathbf{eff},t}$).

the effective helplessness ($h_{eff,t}$), dropping it to near-zero. This allows the base RL policy to resume control and execute the actions needed to escape, providing clear evidence for the function of the antagonistic circuit.

6.3 Sensitivity to Helplessness Threshold θ_h

To understand the impact of our key hyperparameter, we perform a sensitivity analysis by training our agent with different values of the helplessness threshold θ_h. The results in Figure 6 show that performance is robust across a range of values but degrades at the extremes. If θ_h is too low, the agent becomes overly "sensitive," intervening too frequently and disrupting the learning of π_{RL}. If θ_h is too high, the agent becomes "stoic," failing to intervene when needed and succumbing to policy collapse. This demonstrates the critical role of this parameter in balancing standard learning with strategic intervention.

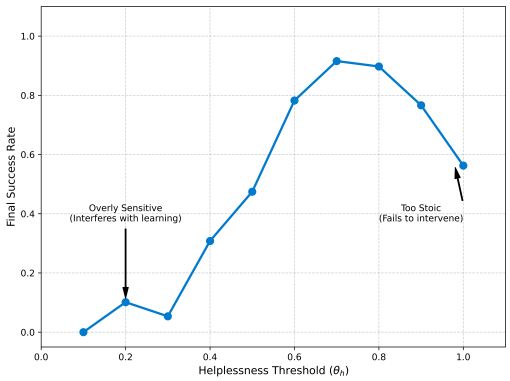

Figure 6: Sensitivity analysis for helplessness threshold (θ_h).

6.4 Generalization to Novel Traps

Figure 7 presents the quantitative results of the zero-shot generalization performance of the trained agents in the novel trap environments. The results strongly support our hypothesis. The AC-DQN agent, despite never having encountered these specific mechanics, successfully generalizes its resilience strategy, achieving a high success rate in both the Sticky Trap and the Teleporter Trap. In contrast, the DQN and DQN+ICM baselines, which failed to solve the original task, are completely unable to cope with the novel traps, achieving 0% success. This demonstrates that the Agency Circuit is not merely overfitting to a specific negative reward signal but is learning a robust, compositional skill: it can identify a general state of helplessness and apply its restorative policy, regardless of the specific cause of that helplessness.

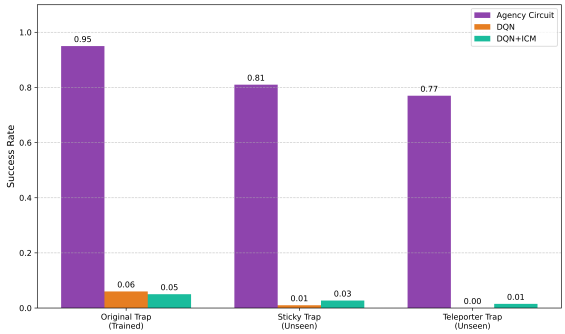

Figure 7: Zero-shot generalization to novel, unseen traps. This bar chart compares the final success rate of all agents across the three environments. The Agency Circuit (AC-DQN) maintains high performance, while baselines fail in all conditions. Success rates are averaged over 10 independent runs.

To provide qualitative evidence, Figure 8 visualizes the AC-DQN agent's trajectories in both environments. These paths show that the agent successfully executes the same fundamental "enter, struggle, escape" behavior in the unseen traps, confirming that it has learned a general, reusable resilience strategy.

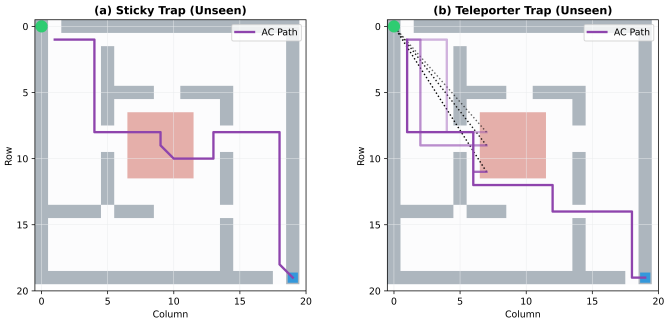

Figure 8: Qualitative generalization on unseen traps. This figure visualizes the successful escape trajectories of the AC-DQN agent. (a) the 'Sticky Trap' and (b) the 'Teleporter Trap'. The trajectories shown are from a single, representative run for each condition.

7 DISCUSSION

Our results demonstrate that the Agency Circuit provides a robust solution to policy collapse in environments designed to induce learned helplessness. The quantitative success of the Agency Circuit, evidenced by its high success rate and low escape latency compared to failing baselines (Figure 3), validates our core hypothesis. More importantly, the analysis of the model's internal dynamics (Figure 5) confirms that this success is achieved through the intended mechanism: an explicit detection of, and targeted intervention against, a state of learned futility. This approach moves beyond generic exploration strategies and offers a principled method for restoring agentic behavior.

The implications of this work extend into several key areas of AI research. First, the Agency Circuit is a step toward more interpretable and explainable RL. The activation of the Control Exertion Module is not an opaque decision from a deep network but a discrete, symbolic event triggered by an observable internal state (h_t). This provides a clear answer to the question of *why* an agent might take a seemingly suboptimal "micro-action" as it is actively working to restore its own sense of control. This aligns with the growing need for transparent AI systems, especially in safety-critical applications.

Second, our findings support the value of psychologically grounded AI. By modeling a well-documented neural mechanism for resilience in mammals, we have created an agent that is not only more robust but also fails in a more understandable way. This suggests that building sophisticated internal state models, reflecting concepts like agency, frustration, or helplessness, is a promising direction for creating agents that can better handle the complexities and setbacks of the real world. The strong zero-shot generalization performance (Figures 7 and 8) further suggests that this resilience mechanism is a robust and reusable skill, not just a solution tailored to a single problem.

Furthermore, it is insightful to consider *why* the CEM's intervention is effective. The success of the "micro-action" does not depend on it being an optimal or task-relevant move. Rather, its purpose is to re-establish the agent's belief in contingency—the fundamental link between action and outcome. This mirrors the psychological principle that overcoming helplessness is less about finding the "right" solution and more about re-learning that solutions are possible at all. While our model is a high-level functional abstraction of the underlying neuroscience, this principle suggests a rich future direction for creating more bioplausible agents that actively manage their own cognitive states.

Lastly, the Symbolic Reasoner relies on classical (bool) logic, using crisp predicates (e.g., $h_t > \theta$) to maximize verifiability. This design choice was deliberate to support interpretability: the CEM triggers because a condition is definitively satisfied. However, extending the reasoner to a non-classical framework (e.g., fuzzy logic) is a compelling direction with clear trade-offs. A fuzzy formulation could enable proportional interventions, potentially improving robustness in highly stochastic or ambiguous environments and smoothing policy behavior. But the cost is reduced explainability: rather than a binary statement ("it acted because IsHelpless is true"), introspection becomes graded ("it acted with magnitude α because helplessness membership was μ"), which may be harder to audit. We believe Boolean logic is appropriate for validating the core mechanism with maximal clarity; future work can explore fuzzy extensions to tune intervention strength once the binary trigger mechanism is fully established.

8 LIMITATIONS AND FUTURE WORK

Our work presents a promising mechanism for agent resilience, but key limitations provide clear avenues for future research. A primary limitation is the reliance on a pre-defined micro-action set, $\mathcal{A}_{\text{micro}}$, which introduces a degree of domain knowledge into the system. Future work could focus on enabling the agent to autonomously discover this set, perhaps by learning a meta-policy to select actions that most effectively restore agency or by identifying actions that are inherently controllable.

Scalability Considerations. Our current demonstration uses a low-dimensional gridworld. Scaling the *Agency Circuit* to pixel-based domains (e.g., *Atari*) mainly requires modifying the *Helplessness Estimator*: replace the one-hot state input with a visual encoder (e.g., CNN) whose features feed the recurrent module (GRU) to compute the history-dependent helplessness signal h_t; the base policy can share the same encoder.

We expect the modular design to transfer because (i) the estimator is trained self-supervised to predict discounted negative return $\hat{G}_t^-$, a signal available in any RL setting, and (ii) the Symbolic Reasoner and CEM operate only on the scalar h_t and simple Boolean predicates (e.g., IsHelpless, RecentlyIntervened), remaining decoupled from raw observations. Our zero-shot results (Figure 7) further suggest the mechanism learns a reusable failure-detection/recovery loop rather than gridworld-specific heuristics.

Practical challenges include: (1) **gradient interference** between the estimator loss and a shared encoder, (2) **long-horizon recurrence** stability (vanishing/exploding gradients), (3) **hyperparameter sensitivity** (e.g., horizon H, loss weights, thresholds), (4) **micro-action design** in large action spaces, and (5) **compute cost** from additional recurrent components.

9 CONCLUSION

In this paper, we addressed the critical problem of policy collapse in Reinforcement Learning by framing it as a computational analogue of learned helplessness. We introduced the **Agency Circuit**, a novel neuro-symbolic architecture inspired by the mammalian brain's resilience circuitry. By integrating a neural state assessor with a symbolic reasoner and a Control Exertion Module, our model successfully detects its own internal state of learned futility and executes targeted interventions to restore its capacity for action. Our experiments demonstrated that this mechanism allows the agent to achieve a high degree of success and efficiently escape deceptive traps where standard and curiosity-driven methods fail completely. Beyond the performance gains, the Agency Circuit enhances agent introspectability by providing a clear, interpretable window into its meta-level decision-making process. This work represents a promising step toward building more robust, explainable, and psychologically-grounded agents. By equipping them with an explicit mechanism for resilience, we move closer to creating autonomous systems that do not just learn what to do, but also possess the intrinsic drive to persevere when learning gets hard.

REFERENCES

[1] Jan Achterhold, Markus Krimmel, and Joerg Stueckler. 2023. Learning temporally extended skills in continuous domains as symbolic actions for planning. In *Conference on Robot Learning*. PMLR, 225–236.

[2] José Amat, Michael V Baratta, Evan Paul, Sondra T Bland, Linda R Watkins, and Steven F Maier. 2005. Medial prefrontal cortex determines how stressor controllability affects behavior and dorsal raphe nucleus. *Nature neuroscience* 8, 3 (2005), 365–371.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[4] Arthur Aubret, Laetitia Matignon, and Salima Hassas. 2019. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976* (2019).

[5] Kate Baumli, David Warde-Farley, Steven Hansen, and Volodymyr Mnih. 2021. Relative variational intrinsic control. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6732–6740.

[6] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems 29*.

[7] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. 2018. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355* (2018).

[8] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894* (2018).

[9] Roberta Calegari, Giovanni Ciatto, Enrico Denti, and Andrea Omicini. 2020. Logic-based technologies for intelligent systems: State of the art and perspectives. *Information* 11, 3 (2020), 167.

[10] George De Ath, Richard M Everson, Alma AM Rahat, and Jonathan E Fieldsend. 2021. Greed is good: Exploration and exploitation trade-offs in Bayesian optimisation. *ACM Transactions on Evolutionary Learning and Optimization* 1, 1 (2021), 1–22.

[11] Shibhansh Dohare, Qingfeng Lan, and A Rupam Mahmood. 2023. Overcoming policy collapse in deep reinforcement learning. In *Sixteenth European Workshop on Reinforcement Learning*.

[12] Artur d'Avila Garcez and Luis C Lamb. 2023. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review* 56, 11 (2023), 12387–12406.

[13] RE Grahn, MJ Will, SE Hammack, S Maswood, MB McQueen, LR Watkins, and SF Maier. 1999. Lesions of the dorsal raphe nucleus block the behavioral and neurochemical effects of inescapable shock. *Brain research* 824, 2 (1999), 223–232.

[14] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. 2016. Variational intrinsic control. *arXiv preprint arXiv:1611.07507* (2016).

[15] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[16] Pascal Hitzler and Md Kamruzzaman Sarker. 2022. Neuro-symbolic artificial intelligence: The state of the art. (2022).

[17] Quentin JM Huys and Peter Dayan. 2016. A computational model of learned helplessness and its reversal by serotonergic agents. *PLoS computational biology* 12, 6 (2016), e1004874.

[18] Steven F Maier. 2015. The dorsal raphe nucleus and the control of defensive behavior: A tale of two streams. *Brain research* 1621 (2015), 192–200.

[19] Steven F Maier and Martin E Seligman. 1976. Learned helplessness: theory and evidence. *Journal of experimental psychology: general* 105, 1 (1976), 3.

[20] Steven F Maier and Linda R Watkins. 2005. Stressor controllability and learned helplessness: the roles of the dorsal raphe nucleus, serotonin, and corticotropin-releasing factor. *Neuroscience & Biobehavioral Reviews* 29, 4-5 (2005), 829–841.

[21] Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* (2018).

[22] Gillian A Matthews, Edward H Nieh, Caitlin M Vander Weele, Sarah A Halbert, Roma V Pradhan, Ariella S Yosafat, Gordon F Glober, Ehsan M Izadmehr, Rain E Thomas, Gabrielle D Lacy, et al. 2016. Dorsal raphe dopamine neurons represent the experience of social isolation. *Cell* 164, 4 (2016), 617–631.

[23] Earl K Miller and Jonathan D Cohen. 2001. An integrative theory of prefrontal cortex function. *Annual review of neuroscience* 24, 1 (2001), 167–202.

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[25] Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. 2017. Count-based exploration with neural density models. In *International conference on machine learning*. PMLR, 2721–2730.

[26] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*. PMLR, 2778–2787.

[27] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. 2019. Self-supervised exploration via disagreement. In *International conference on machine learning*. PMLR, 5062–5071.

[28] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. 2019. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research* 20, 53 (2019), 1–32.

[29] Mohan Raja Pulicharla. 2025. Neurosymbolic AI: Bridging neural networks and symbolic reasoning. *World Journal of Advanced Research and Reviews* 25, 1 (2025), 2351–2373.

[30] Alexander S Rich and Todd M Gureckis. 2018. The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General* 147, 11 (2018), 1553.

[31] Charbel Sakr, Pascal Hitzler, and Amit Sheth. 2022. Neuro-symbolic AI star: A tale of two worlds. *AI Magazine* 43, 4 (2022), 406–419.

[32] Klaus R Scherer. 2022. Learned helplessness revisited: biased evaluation of goals and action potential are major risk factors for emotional disturbance. , 1021–1026 pages.

[33] Martin E. P. Seligman and Steven F. Maier. 1967. Failure to escape traumatic shock. *Journal of Experimental Psychology* 74, 1 (1967), 1–9.

[34] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[35] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[36] Deniz Tuzsus. 2025. *Reinforcement Learning in dynamic environments: Comparing human and artificial recurrent neural networks regarding the exploration/exploitation tradeoff*. Ph.D. Dissertation. Universität zu Köln.

[37] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[38] Jan N Van Rijn and Frank Hutter. 2018. Hyperparameter importance across datasets. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2367–2376.

[39] MR Warden, A Selimbeyoglu, JJ Mirzabekov, M Lo, KR Thompson, S-Y Kim, A Adhikari, KM Tye, LM Frank, and K Deisseroth. 2012. A prefrontal cortex–brainstem neuronal projection that controls response to behavioural challenge. *Nature* 492, 7429 (2012), 428–432.

[40] Michael Wunder, Michael L Littman, and Monica Babes. 2010. Classes of multiagent q-learning dynamics with epsilon-greedy exploration. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (2010), 1167–1174.

[41] Guopeng Zhao, Ailiya, and Zhiqi Shen. 2012. Learning-by-teaching: Designing teachable agents with intrinsic motivation. *Journal of Educational Technology & Society* 15, 4 (2012), 62–74.

[42] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).