Cybernetic Reinforcement Learning Agent for the Wumpus World Environment

Cesar Augusto Pulido Cuervo - 20222020048

Cristian David Romero Gil - 20222020138

Universidad Distrital Francisco José de Caldas

Cybernetic Reinforcement Learning Agent for the Wumpus World Environment

Abstract

Development of an autonomous agent for the Wumpus World, a popular environment in artificial intelligence, can be framed as a dynamical system governed by feedback loops and state transitions. This work integrates cybernetic principles, system dynamics, and AI to create an adaptive agent capable of feedback-based learning in a partially observable and hazardous environment. Key methods include Gymnasium-based simulation and DQN modeling with PyTorch. The system will be evaluated using reward metrics.

1. Introduction

The Wumpus World is a great environment to develop intelligent agents, especially for applying the concepts of cybernetics and the branch of RL in Artificial Intelligence. The approach focuses on the design and improvement of an adaptive agent that can learn to navigate efficiently in the Wumpus World environment. Basically, initially employs a basic Q-learning approach and evolves into a Deep Q-Network framework. Deep reinforcement learning has revolutionized the way artificial agents learn to interact with complex environments. Mnih et al. (2015) introduced the Deep Q-Network algorithm, which allowed an agent to learn directly from visual inputs without the need for feature engineering, achieving human-level performance on various games. Inspired by this approach, this project develops a cybernetic agent for the Wumpus environment, applying deep reinforcement learning techniques to address the perception and decision making challenges that this environment presents.

2. Literature Review

Reinforcement Learning in Grid-Based Environments

Deshpande and Spalanzani (2019) proposed a Deep Q-Network based agent for vehicle navigation, training it in a simulator for a typical intersection crossing setup

amongst pedestrians. Also using reward function, the agent learns a policy capable of driving safely around pedestrians and also follow the traffic rules. As Bora (2024) used in warehouse robot grid-based environment, with a comparative study between two interactive reinforcement learning algorithms: Q-learning and SARSA.

Tošić (2016) propose a functionalist, cybernetics and general systems theory inspired approach to characterizing various types of autonomous agents in terms of their core attributes or capabilities, as evidenced and testable by an external observer.

Wumpus World in Artificial Intelligence

Russell and Norvig (2002) present an approach to artificial intelligence in which agents rely on first-order logic to infer safe actions and deduce hidden state variables. This model marks a shift away from purely reflex-based or hardcoded behaviors, illustrating the power of knowledge-based agents capable of symbolic reasoning under uncertainty. It provides a foundational framework for understanding logical inference and decision-making in early AI systems. Building on this foundation, Kumbhar, Vishwakarma, and Jain (2023) propose a first approximation of our intended agent architecture, also based on first-order logic but extended through the use of the Minimax algorithm.

In their implementation, the environment is modeled as a set of logical propositions, and the agent's knowledge base is composed of declarative rules. Action selection is performed via Minimax search, enabling the agent to evaluate game states and choose actions accordingly. This progression reflects the evolution from theoretical models of logical agents to more structured, decision-oriented implementations. In the other hand, Friesen (n.d.) provides a comparison of exploration techniques for a Q-Learning agent in the Wumpus World, but also with defined environment in the methodology.

As rule-based learning evolves into Deep Q-Network frameworks, the integration of deep learning techniques with decision-making agents has become increasingly relevant. Prior studies on the Wumpus World, reinforcement learning, and cybernetics form the

foundation of this work. Reinforcement learning has shown great promise in addressing complex decision-making tasks, while cybernetic principles—such as feedback loops and adaptive control—are increasingly applied in the design of intelligent systems.

3. Background

The Wumpus World consists of a grid-based environment populated with hazards such as pits and the Wumpus creature, as well as a pieces of gold that the agent is tasked with retrieving. The environment is partially observable: the agent does not have full knowledge of the map and must infer danger zones based on limited local percepts—breeze (indicating nearby pits), stench (indicating proximity to the Wumpus), and glitter (signaling the presence of gold). These percepts form the foundation of the agent's sensory input and are updated at each time step. A feedback-driven agent design integrates three core components: percept interpretation, action selection, and environment response. The agent processes incoming sensory data, updates its internal state or knowledge base accordingly, selects an appropriate action, and then observes the consequences of that action, forming a continuous perception–action loop. This architecture exemplifies the principles of deliberative reasoning and learning under uncertainty, making the Wumpus World a canonical testbed for knowledge-based agents and a valuable abstraction for more complex AI systems.

4. Objectives

- Design and implement an intelligent agent capable of operating within a partially observable, dangerous environment such as the Wumpus World, using reinforcement learning principles and adaptive control mechanisms.
- Integrate cybernetic feedback principles into the agent's decision-making loop, allowing for self-regulation and behavioral adaptation through perception–action cycles grounded in system dynamics.

- Incorporate a Deep Q-network architecture that allows for nonlinear approximation of value functions, enabling the agent to generalize across diverse environment states and learn stable policies.

- Evaluate agent performance using multi-dimensional metrics, including cumulative reward, step efficiency, action entropy, and task success rate under various environmental conditions and random seeds.

5. Scope

The study is bounded by several exclusions. It does not explore continuous action spaces, symbolic logic-based reasoning, or transfer learning mechanisms. The system operates exclusively in simulation and does not include real-world deployment, robotic embodiment, or hardware interfacing. Additionally, the learning architecture is restricted to Q-learning and Deep Q-Networks without incorporating more advanced extensions such as Double DQN, Dueling DQN, or memory-augmented agents.

These boundaries are established to maintain a focused investigation into feedback-based reinforcement learning using cybernetic principles and to ensure experimental reproducibility within a well-defined and interpretable environment.

6. Assumptions

The following assumptions were made during the development and evaluation of the reinforcement learning agent in the Wumpus World environment. These assumptions help establish a controlled context for experimentation but may also limit the generalization of the findings:

- **Accurate and Noise-Free Percepts:** The perceptual signals received by the agent are assumed to be fully accurate and deterministic. No perceptual noise or uncertainty is introduced in the observation model.

- **Fixed Initial Conditions:** The agent always begins from the same starting position, ensuring uniformity across training sessions and simplifying policy evaluation.
- **Closed Simulation Environment:** All interactions occur within a simulated grid world implemented in Gymnasium, without uncertainties such as hardware limitations or sensor inaccuracies.
- **Independent Episodes:** Each episode is treated as an independent trial with no memory or transfer of experience between episodes.

7. Limitations

In addition to the methodological constraints defined in the project scope, this research was influenced by several practical and contextual limitations that may have affected the depth, efficiency, and generalization of the results.

- **Time Constraints:** The project was conducted over a period of approximately 14 weeks, which limited the extent of experimentation, hyperparameter tuning, and long-term testing. As a result, certain design iterations—such as architecture optimization or ablation studies—could not be fully explored.
- **Technical Learning Curve:** The implementation required the acquisition of new knowledge in areas such as deep reinforcement learning, PyTorch programming, Gymnasium environment design, and system modeling. Time spent on self-learning and troubleshooting reduced the capacity for extensive experimentation and model refinement.
- **Limited Testing Infrastructure:** The agent was trained and tested on standard local hardware without access to GPUs or distributed training environments. This limited the number of training episodes that could be completed and restricted the scale of the neural network architecture.

- **Simplified Evaluation Metrics:** Due to time and tooling constraints, evaluation focused on a core set of metrics (reward, success rate, entropy) without more advanced diagnostic tools such as policy visualization, saliency maps, or in-depth behavioral profiling.

- **No External Benchmarks:** The project did not incorporate or compare against publicly available Wumpus World benchmarks or alternative agents, which restricts the ability to position the results within a broader research context.

- **Manual Debugging and Testing Biases:** Some evaluations and observations relied on manual inspection (e.g., visual trajectory analysis), which may introduce subjectivity or overlook edge-case behaviors.

Acknowledging these limitations provides a more accurate interpretation of the project outcomes. While the results are promising within the controlled simulation environment, future iterations with extended time, computing resources, and methodological refinements would be necessary to validate the system's robustness and broader applicability.

8. Methodology

This work adopts a modular and simulation-driven methodology for the design, training, and evaluation of a reinforcement learning agent operating in a Wumpus World environment. The environment was developed using the Gymnasium framework, which provides a standardized interface for defining observation and action spaces, episodic structure, and reward dynamics (Terry et al., 2023). The environment encodes partial observability through discrete percepts—breeze, stench, and glitter—mapped to an agent-centered coordinate system. To monitor the simulation visually and assist in debugging, a rendering layer was implemented using Pygame.

The agent architecture is based on DQN implemented in PyTorch. This neural model

receives an encoded state vector representing the current perceptual inputs, positional data, and a computed danger level. The network is trained using experience replay and stabilized via the use of target networks, in accordance with best practices in deep reinforcement learning. Each episode is structured as a series of decision steps, during which the agent selects actions using an ϵ-greedy policy that decays over time to encourage initial exploration followed by exploitation.

A key component of the training process is the use of rewards. The agent receives feedback such as:

Event	**Reward**
Falling into a pit	-1000
Being caught by Wumpus	-1000
Unsuccessful shooting	-25
Step	-5
Revisiting safe cell	-0.5
Visiting a new safe cell	+5
Passing near danger and survive	+10
Collecting gold	+1000
Winning the episode	+2000

This feedback mechanism serves to guide learning more efficiently by embedding behavioral incentives aligned with safe exploration and task completion. The high-level architecture of the system is clearly depicted in the figure.

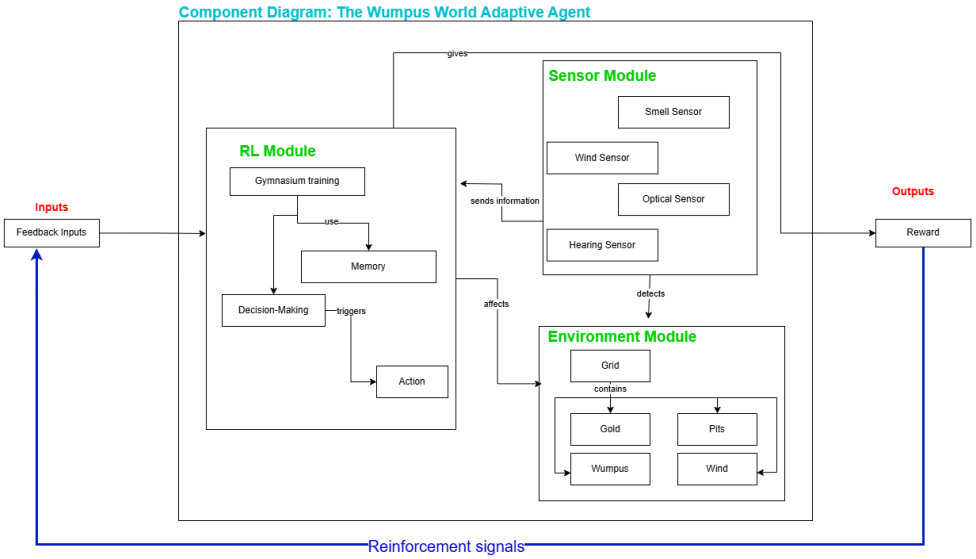

Figure 1. High-level architecture illustrating the integration of three main modules

Figure 1 depicts the high-level architecture of the simulation system, highlighting the interaction between environment, agent, and user interface components. The design supports modular experimentation and enabling updates to reward structures,

Incorporating feedback mechanisms into the learning loop is crucial for achieving convergence in dynamic and partially observable environments. Figure 2 presents the closed-loop control system, wherein perceptual input triggers action selection, which in turn updates the environment and generates new observations. This design enables learning from the consequences of prior decisions, allowing the agent to refine its policy through iterative trial and error.

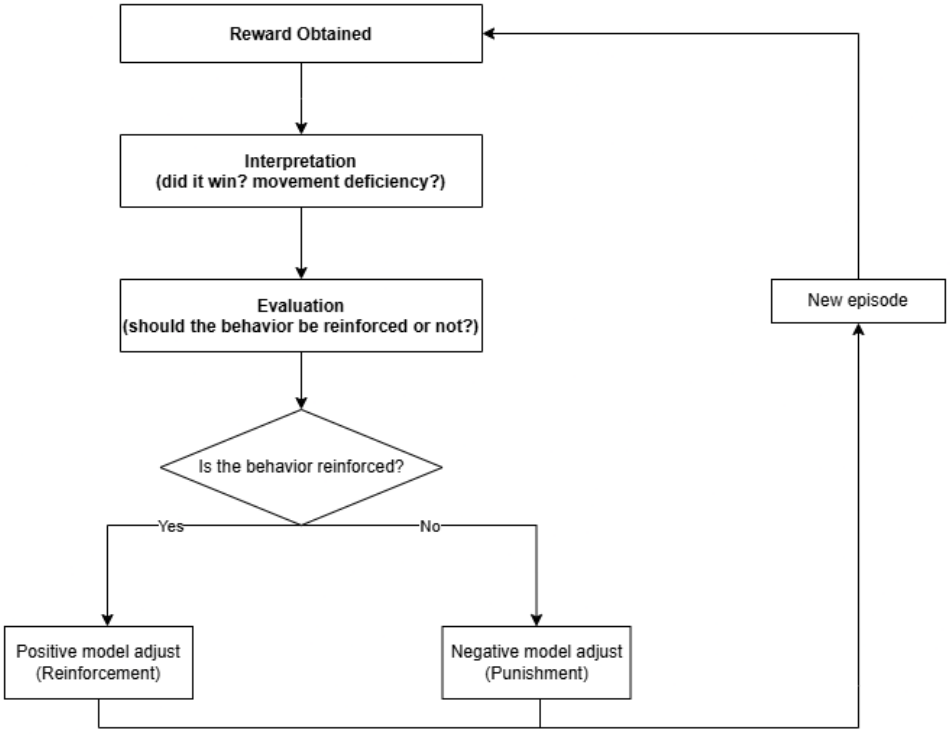

Figure 2. Feedback loops illustration

Additionally, the Wumpus World domain was modeled using a causal diagram to represent the interdependencies among environment variables, percepts and the dynamical behavior of the system. As shown in Figure 3, each sensory input is causally linked to hidden state elements such as nearby pits or the presence of the Wumpus.

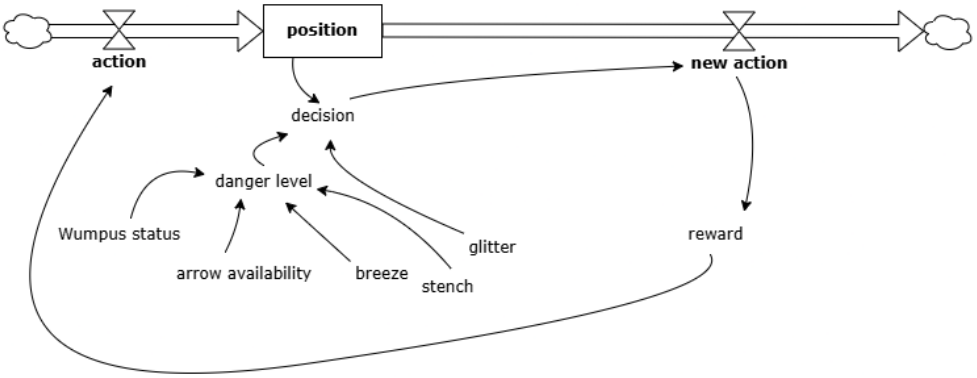

Figure 3. Causal diagram

9. Conclusion

The agent leverages a DQN architecture enhanced with reward shaping and perceptual modeling to improve learning efficiency in a partially observable, hazardous environment. The use of a custom Gymnasium-based environment, along with real-time visualization through Pygame, supports both interpretability and reproducibility.

Although full implementation is still in progress, the system design emphasizes modularity, interpretability, and convergence through structured testing strategies. Key contributions include the introduction of a continuous danger-level sensor, a feedback-informed reward scheme, and an architectural feedback loop inspired by cybernetic control systems.

The implications of this work extend to both AI education and practical agent design. The simulation serves as a pedagogical tool for understanding learning under uncertainty and provides a testbed for evaluating adaptive behavior in constrained environments.

References

Bora, A. (2024). A comparative analysis of interactive reinforcement learning algorithms in warehouse robot grid based environment. *arXiv preprint arXiv:2407.11671*.

Deshpande, N., & Spalanzani, A. (2019). Deep reinforcement learning based vehicle navigation amongst pedestrians using a grid-based state representation. In *2019 ieee intelligent transportation systems conference (itsc)* (pp. 2081–2086).

Friesen, A. (n.d.). A comparison of exploration/exploitation techniques for a q-learning agent in the wumpus world.

Kumbhar, M., Vishwakarma, V., & Jain, R. (2023). A logical agent approach to solving the wumpus world problem: An analysis of game trees. In *2023 9th international conference on advanced computing and communication systems (icaccs)* (Vol. 1, pp. 1839–1844).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Russell, S. J., & Norvig, P. (2002). A modern approach. *Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems*, *90*, 33–48.

Terry, J., Chan, K., Tucker, G., Zholus, A., Gao, X. B., Agarwal, R., et al. (2023). *Gymnasium: A standard api for reinforcement learning environments.* `https://gymnasium.farama.org`. (Farama Foundation)

Tošić, P. T. (2016). Understanding autonomous agents: A cybernetics and systems science perspective. In *2016 future technologies conference (ftc)* (pp. 121–129).