Maze Pirates.

Eric L. Foster

SNHU CS-370

10/17/2025

A deep Q-learning intelligent agent was developed in this project and was subsequently trained to navigate an agent masquerading as a pirate in a treasure hunt maze. The pirate had its objectives defined as navigating obstacles and locating the treasure before the player. The maze was a pathfinding environment in which the agent received rewards for successful moves and penalties for collisions or dead ends. The assignment was to apply reinforcement learning principles, specifically the Q-learning algorithm enhanced with a neural network, to teach the agent to make sequential decisions that maximize cumulative reward over multiple episodes.

A human player would typically rely on a combination of perception, planning, reasoning, and frustration to solve a maze. Such a person might visually scan the maze, choose a direction that hopefully leads toward the goal, remember past turns, and avoid revisiting dead ends. Leaning heavily into Theory of Mind, humans make use of cognitive mapping and heuristic shortcuts that rely on prior experiences and intuition (Kahneman, 2011). They can generalize patterns, such as recognizing symmetrical structures or identifying paths that lead back to previously visited points. As they explore, humans balance trial and error with logical deduction, building a mental representation of the environment that evolves with each decision (Lindsay, 2021).

In contrast, the intelligent agent does not rely on reasoning or visual cues but instead learns through experience and feedback from the environment to which it is subjected. The pirate agent perceives its current state as a vector of numerical values, takes one of the four possible actions (moving up, down, left, or right), and receives a reward signal. With each experience consisting of a state, action, reward, next state, and completion flag being stored in memory. During training, the agent samples batches of these past experiences to update a neural network that approximates the state–action value function. Over repeated iterations, the pirate learns which actions in a given state are most likely to lead to the treasure. This process mirrors how humans learn from repeated practice, but instead of abstract reasoning, the agent’s learning occurs through statistical approximation and iterative updates of network weights.

Both human and machine approaches rely on incremental learning and feedback. However, humans adapt quickly and can generalize knowledge between different maze structures, while the agent must experience each configuration repeatedly to learn effective behavior. Humans on a whim recognize spatial relationships and can use short-term memory or deduction to plan several moves ahead (Mohammed Ali & Gonzalez, 2023). The agent lacks such cognitive structure and instead depends on a reward-driven optimization process to discover the correct sequence of actions.

The purpose of the intelligent agent is to autonomously learn a policy that efficiently finds the treasure. In reinforcement learning, this involves balancing two competing objectives: exploration and exploitation (Yan et al., 2024). Exploration allows the agent to try new actions to gather information about unknown paths, while exploitation leverages current knowledge to maximize rewards. In the implemented algorithm, an ε-greedy exploration strategy was used, beginning with a high exploration rate and gradually decaying it as training progressed (Miyashita & Sugawara, 2022). In early episodes, this allowed the pirate to discover the layout of the maze, while in later stages, lower exploration promoted consistency and refinement. When the decay rate was too high, the agent began exploiting too early and became trapped in suboptimal paths, as seen in Experiment B. This demonstrated that maintaining exploration for a sufficient period is essential for robust learning.

The deep Q-learning algorithm used in this project applies a neural network to approximate the Q-function instead of having to rely on a lookup table (Perić et al., 2022). The model consists of multiple dense layers with ReLU activations and a linear output layer representing the Q-values for each possible action. The network predicts the expected cumulative reward from any given state–action pair and updates its weights through backpropagation. Experience replay was implemented to improve stability and efficiency by sampling random batches of past experiences, preventing the model from overfitting to recent transitions. A target network was also used to stabilize training by providing consistent Q-value estimates during updates. Together, these mechanisms allow the agent to generalize across similar states and gradually converge toward an optimal navigation strategy.

The implementation of deep Q-learning in this game required careful tuning of parameters such as learning rate, discount factor, and exploration decay. These hyperparameters directly affected model performance. A higher learning rate enabled faster convergence but risked overshooting optimal weights, while a lower rate produced smoother and more stable progress. The discount factor controlled how strongly the agent valued future rewards, encouraging prioritization of long-term success (reaching the treasure) over immediate gains. With these components integrated effectively, the final agent achieved functional and reliable navigation behavior without runtime errors, this project effectively demonstrates how reinforcement learning and neural networks can be combined to solve complex pathfinding problems in an interactive game environment.

References

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux. <https://doi.org/10.1037/e593572012-001>

Lindsay, G. (2021). *Models of the mind: How physics, engineering and mathematics have shaped our understanding of the brain*. Bloomsbury Sigma. <https://doi.org/10.5040/9781472966445>

Miyashita, Y., & Sugawara, T. (2022). Flexible Exploration Strategies in Multi-Agent Reinforcement Learning for Instability by Mutual Learning. *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Machine Learning and Applications (ICMLA), 2022 21st IEEE International Conference on, ICMLA*, 579–584. <https://doi.org/10.1109/ICMLA55696.2022.00100>

Mohammed Ali, A. E., & Gonzalez, A. J. (2023). Machine learning from casual conversation. *Machine Learning*, *112*(12), 4789–4836. <https://doi.org/10.1007/s10994-023-06383-0>

Yan, R., Gan, Y., Wu, Y., Liang, L., Xing, J., Cai, Y., & Huang, R. (2024). *The Exploration-Exploitation Dilemma Revisited: An Entropy Perspective*.

Zoran Perić, Aleksandar Marković, Nataša Kontrec, Jelena Nikolić, Marko D. Petković, & Aleksandra Jovanović. (2022). Two Interval Upper-Bound Q-Function Approximations with Applications. *Mathematics*, *10*(19), 3590. <https://doi.org/10.3390/math10193590>