Arielle Moore

CS 370: Current/Emerging Trends in CS

February 16th, 2026

# Treasure Hunt Game: Design Defense

Complex problems, such as that of navigating a maze, are approached by humans and machines with fundamentally different processes. A human endeavoring to solve a maze will typically begin by visually scanning the environment to identify features like open paths, obstacles, and dead ends (Cherry, 2025). The human player uses a combination of spatial reasoning, working memory, and heuristic strategies, often recalling previously-visited locations to plan the next move and avoid unnecessary backtracking (Cherry, 2025). Conscious reasoning coalesces with trial and error to learn from past mistakes, allowing the human player to continuously refine strategies (Cherry, 2025). In contrast to the human process, the intelligent agent (as exemplified in our game) solves the maze using deep reinforcement learning, thereby representing the environment numerically rather than visually (Sutton & Barto, 2020). The agent encodes the pirate's position and maze layout into a flattened vector and selects actions based on predicted Q-values from a neural network (Sutton & Barto, 2020). Feedback is then represented in the forms of rewards and penalties, which allow the agent to iteratively update its policy. A human player employs their conscious reasoning whereas the agent relies on statistical patterns learned from experience; however, both the human player and the agent adapt based on feedback and experience (Cherry, 2025; Sutton & Barto, 2020).

In our Treasure Hunt Game, the agent's purpose is to autonomously discover an optimal path to the treasure while maximizing cumulative rewards. With this, a key

consideration in reinforcement learning is achieving a harmonious balance between exploration and exploitation. *Exploration* involves testing unfamiliar actions to gather information about the environment, whereas *exploitation* selects actions known to yield high rewards (Sutton & Barto, 2020). Striking this balance ensures that the agent does not prematurely converge on suboptimal paths and instead learns effective strategies over time (Sutton & Barto, 2020). In our maze project, an ε-greedy policy was implemented with ε decaying over training (Sutton & Barto, 2020); thus, a slightly higher emphasis on exploration at the beginning followed by a gradual shift to exploitation provided the ideal balance to efficiently discover the treasure while avoiding dead ends (Sutton & Barto, 2020). The agent initially explored widely to build a robust understanding of the maze at hand, then gradually shifted to exploitation as its learning progressed (Sutton & Barto, 2020). This approach directly mirrors the human strategy of experimenting in unfamiliar areas before it is able to rely on learned shortcuts (Cherry, 2025).

Using a reward-based system, reinforcement learning enables the agent to evaluate the consequences of actions (Sutton & Barto, 2020). Our maze game features small negative rewards meant to discourage inefficient movement, whereas reaching the desired treasure yields a positive reward (Sutton & Barto, 2020). Invalid or blocked moves incur higher penalties, and revisiting cells is discouraged through incremental negative rewards (Sutton & Barto, 2020). Experiences such as state, action, reward, next state, and terminal status are all stored in the agent's memory, which are then used as sample batches in order to update the Q-function (Sutton & Barto, 2020). This is known as *experience replay*, a technique that reduces correlation between sequential experiences and stabilizes learning (Sutton & Barto, 2020). The neural network approximates the expected cumulative reward for each action in any state over time, resulting in the agent's ability to generalize across the environment and make optimal decisions (Sutton & Barto, 2020).

The deep Q-learning implementation uses a neural network that maps the flattened maze state to Q-values for all possible actions (Sutton & Barto, 2020). The process of each training iteration consists of taking an action via the ε-greedy policy, storing the resulting experience, sampling batches from memory, and updating the network using the *Bellman equation* (which provides a recursive relationship to calculate the expected cumulative reward for each state-action pair, guiding the agent's Q-value updates), allowing the agent to iteratively improve its policy to achieve higher rewards (Sutton & Barto, 2020). We can observe from our training epochs that the agent's success rate steadily increased, thus demonstrating the overall effectiveness of reinforcement learning for autonomous pathfinding (Sutton & Barto, 2020).

In conclusion, the agent exhibits a learning process conceptually similar to that of the human player in its iterative adaptation to feedback, yet still fundamentally different in both representation and reasoning (Cherry, 2025; Sutton & Barto, 2020). Humans rely on conscious spatial reasoning and heuristics (Cherry, 2025), whereas the agent uses numerical state representations and reward-driven updates (Sutton & Barto, 2020). The agent efficiently balances exploration and exploitation by employing deep Q-learning and experience replay, which ultimately leads to the discovery of optimal paths to the goal (Sutton & Barto, 2020). The design of the Treasure Hunt Game appropriately demonstrates the robustness of reinforcement learning in complex decision-making tasks.

## References

Cherry, K. (2025, November 13). *What Is Confirmation Bias?* Verywell Mind. https://www.verywellmind.com/what-is-confirmation-bias-2795024

Sutton, R. S., & Barto, A. G. (2020). *Reinforcement Learning: An Introduction* (2nd ed.). http://incompleteideas.net/book/RLbook2020.pdf