**7-3 Project Two: Design Defense**

Southern New Hampshire University

CS 370 Current/Emerging Trends in CS

Chris Bridges

Professor Oakes

19 October, 2025

There are a few key differences in how a human and a machine solve this maze. I learned as a firefighter to perform a “left-handed Search”, where a maze was completed while maintaining contact with the left wall and completing a maze. It was far from a perfect approach, but it worked well with no visual input, although it worked very poorly in large open areas. Assuming no visual impairment, a human would typically identify the start and end points, obstacles, and potential paths. The human would then move towards the goal. If a dead end was encountered, they would return to the last junction and try an alternate route. Our agent knew only the cell it was in, made a move, and received feedback. Using this feedback, it would remember the optimal route to take. It would explore in the beginning, finding the best path, then use exploitation to run the optimal path numerous times. Both the human and machine methods learn from the past, find the best route, and take it. Humans, however, can generalize quickly and memorize the maze in one to two runs, while a machine takes many more trials before gaining confidence. Another difference is intent and awareness. A human understands why they are navigating the maze, to find a treasure, navigate a maze, or exit a building. The agent has no concept of purpose. It doesn’t want to solve the puzzle; it just follows mathematical updates that reward its behavior. A human could get tired, bored, upset, or frustrated in a maze, while an agent would never get distracted or careless. Once the agent is trained, it will find the most optimal path and take it every time without hesitation.

The agent I trained had no understanding of the maze layout. It knew only the single cell it occupied at each step. It did not see the goal or remember landmarks; instead, it took an action and received feedback in the form of a reward or penalty. Through repeated trials, it gradually learned which moves led to progress and which resulted in failure. Early in training, it relied on exploration, testing many possible paths, even bad ones. Over time, as it discovered a successful route, it shifted into exploitation, repeating what worked. Both humans and machines ultimately learn from past experiences and focus on efficient routes. Still, humans can infer patterns and generalize after one or two attempts, while the agent requires hundreds of episodes to build statistical confidence. The purpose of the agent is to learn a policy. Map from states to actions that maximize reward. Exploration is like wandering unfamiliar hallways to see what exists, while exploitation is like confidently using a known shortcut once it’s proven to work. The ideal strategy, especially for this maze, is to start with heavy exploration and gradually reduce it to a low, stable value like ε\_min ≈ 0.05 (Sutton & Barto, 2015), allowing the agent to refine what it has learned rather than constantly guessing. Reinforcement learning applies the Bellman Equation, repeatedly updating the value of actions based on future rewards. This process biases the agent toward actions that lead to the treasure quickly and discourages wandering or revisiting dead ends.

I used Deep Q-Learning (DQN) to train the pirate to solve the maze. I built a simple neural network that looks at the maze (as a flat list of numbers) and predicts which move is best, using Left, Right, Up, or Down. After each move, the agent gets a reward or penalty. It learns to choose actions that lead to higher rewards, especially moves that eventually reach the treasure. The agent stores past moves in memory and replays them later to learn from them again. This prevents it from forgetting good strategies. At first, the agent uses exploration, trying random moves. Over time, it switches to exploitation, where it uses what it has learned. I slowly reduced randomness, so it focused on winning paths. A second copy of the network, the target network, is updated every few episodes (Mnih et al, 2015). This keeps training stable and prevents wild swings in learning and plateaus. Training stops when the agent consistently wins with nearly 100% success across all starting positions.

**References**

Mnih, V., Kavukcuoglu, K., Silver, D. *et al.* Human-level control through deep reinforcement learning. *Nature* 518, 529–533 (2015). <https://doi.org/10.1038/nature14236>

Sutton, R. S., & Barto, A. G. (2015). *I reinforcement learning: An introduction second edition, in progress*. Reinforcement Learning: An Introduction. https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf