**Jerry Vasquez**

**CS 370**

**Design Defense**

A human being would attempt to solve this maze by keeping track of which way they are going as they choose a path. A human would most likely pick a path by random and make assessments as they explore the path. When a human picks a path that leads to a dead end, they are likely to back track to the intersection where they initially made the turn and pick a different path. At that point, the human would remember that the previous path leads to a dead end and would avoid going down that path again. The human can continue this approach of exploring new pathways, back tracking, and eliminating pathways that lead nowhere until they find the right path to the treasure.

To find the best path to the treasure, the intelligent agent uses reinforcement learning in the form of a deep Q-learning algorithm. The agent first observes the current state of the environment, the maze, and identifies its possible actions (up, down, left, right). It then chooses an action based on either an exploration (trying new moves) or exploitation (using what it has already learned) approach. After moving, the agent records its experience, including the new state, the reward received, and whether the game has ended. The agent tries to find a balance between exploration and exploitation to solve the maze efficiently (Murel, n.d). Over time, as the agent collects and stores many experiences, it uses this data to predict the next best possible actions that maximize rewards (Murel, n.d). Through repeated simulations, the agent gradually learns the best path to reach the treasure.

The similarities between both approaches are that the human and intelligent agent both use their past experiences to influence their decisions. They both also have instances where they choose a path at random. A key difference is how each one stores their past experience. The intelligent agent stores information such as previous state, action, and rewards in a data structure from which they can easily reaccess, while the human needs to just go off of memory. This can be a limitation for the human since humans can forget which path they have already taken making them disoriented.

The intelligent agent’s decision making strategy can follow two different approaches which are exploitation and exploration. Exploration aims to gain new information about the environment by selecting actions where the results such as state and rewards may be unknown (GeeksForFeeks, 2025). On the other hand, the exploitation approach opts to use previously gained information to select an action that is more likely to be favorable or maximize rewards (GeeksForFeeks, 2025). An ideal decision making strategy would find the right balance between exploitation and exploration (GeeksForFeeks, 2025). The ideal proportion of exploration and exploitation can be achieved for this maze problem with the epsilon-greedy strategy that utilizes an epsilon decay factor.

Reinforcement learning excels at achieving optimal outcomes in complex and unfamiliar environments such as this maze (AWS, n.d). This makes reinforcement learning a suitable method for determining the correct path to the treasure by allowing the pirate to efficiently learn its environment and the best actions at a particular state (Murel, n.d). By introducing rewards for positive actions and finding the treasure faster, the pirate is incentivised to find the optimal path (AWS, n.d).

For this game, we implement the Deep Q-Learning algorithm by starting with representing the maze as an 8x8 matrix where the 0s represent an obstacle and 1s represent an open tile. This way the agent can track its current position and next available moves. The pirate starts at the top left corner of the maze and observes the current state of the environment. The pirate will then decide on an action based on an epsilon-greedy policy where the pirate will either choose a random action (exploration) or pick the next best move based on its previous experiences (exploitation). We select a random number between 0 and 1 where if the random number is smaller than epsilon, then the pirate opts for exploration, else, the pirate opts for exploitation. For exploitation, the neural network uses a random batch sample of past experiences as training data to predict the Q-values of each next available action and choose the action with the highest Q-value, the next best move. We opt for a random batch sample for training the neural network to stabilize learning and prevent the network from being biased towards more recent actions (DeepChecks, n.d). The pirate then performs the action and records information in its experience replay memory such as the action made, previous and current state, reward, and status of the game. The pirate will continue this process until the game is over and record the win or loss, completing one episode or full run of the game. We continue to run various episodes until the win rate of the pirate is 95% or higher.

**References**

* GeeksForGeeks. (2025, July 23). Exploitation and Exploration in Machine Learning. GeeksForGeeks. <https://www.geeksforgeeks.org/machine-learning/exploitation-and-exploration-in-machine-learning/>
* Murel J. (n.d). What is reinforcement learning?. IBM. <https://www.ibm.com/think/topics/reinforcement-learning>
* AWS (n.d). What is Reinforcement Learning?. AWS. <https://aws.amazon.com/what-is/reinforcement-learning/>
* DeepChecks. (n.d). Deep Q-Network. DeepChecks. <https://www.deepchecks.com/glossary/deep-q-network/>