Apurva Shukla

CS-370

16th April 2023

**Design Defense**

The process to solve a maze is not very different in humans. It always follows a similar pattern to pathfinding, such as trying out different routes or even just perusing the maze with the intent to solve it on the first try. Trial and error is the most popular way of solving mazes, and this is the method that most humans will use to solve mazes. The intelligent agent, or the pirate in this case, uses a similar form of maze solving, but instead using a combination of both pathfinding and random exploration. The agent then saves the state of the current decision that it has taken and uses it for the next move to determine whether to move to that next state or not. These decisions are aided by a system of rewards that have been given to the agent.

Exploitation in terms of decision selection for AI is taken to be the most optimal decision with respect to the data observed. This approach is considered safe and tries to avoid bad decisions. Exploration in terms of selection consists of not taking decisions that seem to be optimal and instead taking ones that observed data is not sufficient to identify. According to the article written by Joseph Rocca, “The problem of choosing between exploitation and exploration can be encountered in many situations where observations drive decisions and decisions lead to new observations.” This is an important distinction to make as there is no perfect way of knowing if the given data is sufficient enough to train the model properly and is a common problem that is faced by engineers who train these models. The article mentions four different methods to tackle this problem, which are e-greedy methods, optimistic initialization, upper confidence bounds and Thompson sampling. From these methods, the method that seems to give the most ideal proportion is optimistic initialization. In optimistic initialization, “…although they are often more complex, there exist other strategies that make possible to resolve uncertainty in a smarter way by focusing the exploration on the most relevant actions that could still be proven to be optimal (Rocca, 2022)”. By using this method, we can tell the model that the actions that should be picked can be assumed to be the best for expected rewards until it is proven otherwise by the data or by the model’s own learning, basically translating into experience more than random luck and pure by-the-books following that the model does by only following one of the two. Reinforcement learning here will help the agent to find the treasure in the most efficient and quickest way possible, as it will save all the more rewarding parts in it’s learning model.

The deep Q-learning algorithm implemented was by giving the model certain set of data which allowed it to understand the paths needed to take to complete the maze, and variables to allow it to store the different states and actions that the algorithm will decide to take and putting out a success message when the treasure has been found.

**References**

Rocca, J. (2022, January 7). The exploration-exploitation trade-off: intuitions and strategies | by Joseph Rocca | Towards Data Science. *Medium*. https://towardsdatascience.com/the-exploration-exploitation-dilemma-f5622fbe1e82