## Results:

#### Simple Push Environment

Both the PPO and the DQN algorithms showed good learning results in the simple push environment as can be seen from the charts below:

Chart, scatter chart

Description automatically generated

Figure : learning curve from simple push environment using the DQN algorithm

Chart, line chart

Description automatically generated

Figure : learning curve from simple push environment using the PPO algorithm

Both show convergence to an optimal policy for both the agent and adversary, with the DQN reaching a seemingly optimal policy for the adversary within 150 episodes, while the PPO seems to take around 600 episodes for both agent and adversary to converge.

#### Simple World Comm Environment

The same algorithms were then applied to the simple world comm environment. In this the PPO algorithm performed better with the learning curves for both the agent and the adversary showing good learning during the initial ~100 episodes and convergence to a seemingly optimal policy by ~250 episodes.

Chart, line chart

Description automatically generated

Figure : Learning curves for Good Agent and Adversary over 1050 episodes using the PPO algorithm

Chart, line chart

Description automatically generated

Figure : Learning curves for Good Agent and Adversary over 500 episodes using the DQN algorithm

Chart, bar chart

Description automatically generated

Figure : Learning curve showing possible strategy development of agent and adversary

The other main feature of the learning curves for the PPO algorithm is when the comparison of rewards for the agent and adversary are compared once the initial learning phase has been completed. As we can see from the chart above, there is a clear oscillation of the reward curves for the agent and adversary. These responses have been observed in similar MARL environments (Baker et al, 2020), and represent the development of a new strategy and then this being countered and overcome by the opposing agents/adversary. This strategy development seems to have at least five phases during the training and have been highlighted in the chart the reflect the agent or adversary that appears to have an advantage during these episodes.