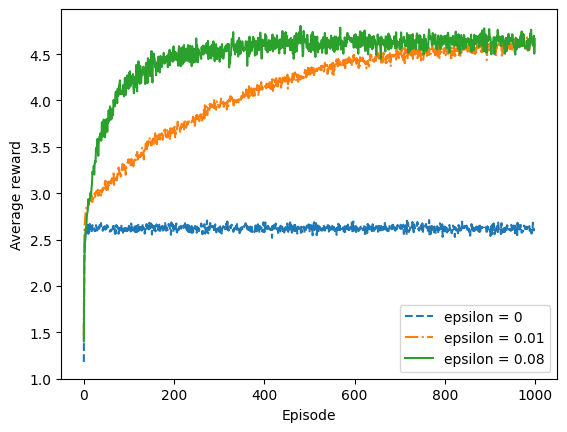
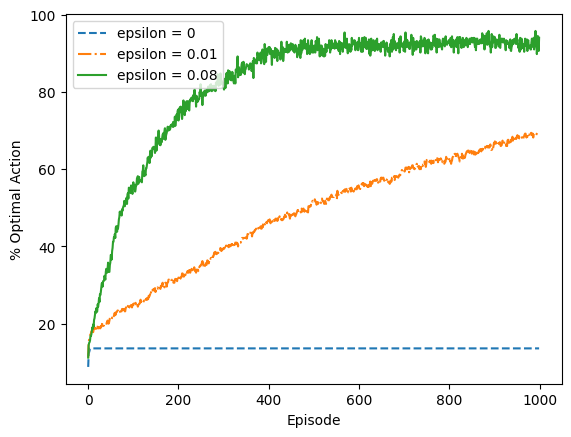
The effect of the epsilon parameter was evaluated in multiple environments. More specifically, we seeded three gambling environments with values 5, 1000, and 10,000 respectively. In each environment, we ran a simulation of 1,000 runs and 1,000 episodes and with epsilon values of 0, 0.01, 0.03, 0.05, 0.08, 0.1, and 0.15 respectively. Each combination of epsilon and seed value was evaluated based on the average reward per episode and percentage of optimal action produced during the simulation. In addition, the degree to which the seeding affected the action counts was also considered to measure how exploratory the combination tended to be. The results are shown in section **5.2** of the code.

Figure 1 illustrates the effects of different values of epsilon on the average reward and the percentage of optimal action taken, per episode. An epsilon value of 0.0 corresponds to a greedy algorithm that only selects the action with the highest Q-value. This algorithm lacks an exploratory part, which often causes it to remain stuck choosing suboptimal actions, resulting in lower average rewards. An epsilon value of 0.01 explores 1% of the time and is therefore able to find the optimal action more often, also resulting in higher average rewards. However, after 1,000 episodes, the optimal action percentage and average rewards have not had enough time to converge and fully learn the optimal action. An epsilon value of 0.08 converges to a percentage of 91% optimal action after circa 500 episodes. As a result, the average reward is highest for this epsilon value. If episodes were longer, it is expected that for an epsilon value of 0.01, both the average reward and percentage of optimal action exceed that of the epsilon value of 0.08.





**Figure 1**: Average performance of epsilon-greedy action-value simulations. The data is averaged over 500 runs.