My research experiences at Georgia Tech have been bountiful, but that which I found most stimulating are those experienced in the realm of reinforcement learning. I have done two major reinforcement learning projects. The first, a moon lander simulation, wherein the agent attempts to use left, right, and bottom thrusters to safely land on a level pad on the otherwise craterous moonscape. The second, a soccer environment with multiple agents working together to achieve better results than any one agent can alone.

The moon lander project was an exploration of discretizing a continuous observation space via neural network, and otherwise solving the system as a markov decision process, letting the neural network take the place of Bellman's Q-function. Additionally, concepts like memory replay were used to introduce stability into the agent's learning process (courtesy of Mnih et. al, 2015). This project was of particular challenge for me as a bug gone undiscovered provided for less than ideal results. This prompted me to revisit the project between semesters, again and again, providing countless hours of experience in hyperparameter tweaking, including benefits and drawbacks of hyperparameter schedules. Further, no longer under pressure of a due date, I was able to broaden the scope of the project and got to explore parallel training including the implementations and configurations relevant to sharing between multiple CPU cores and the GPU.

The soccer project was an exploration of multi-agent environments. For this, I performed an ablation study comparing the Proximal Policy Optimization algorithm to several algorithm updates. These changes are implemented with the intent to introduce agents more aware of their teammates and adversaries. The research involved in this project lead to some very interesting strategies including the memory intensive Multi-Agent Deep Deterministic Policy Gradient (MADDPG, Lowe et. al. 2017), and the more accessible Counterfactual Multi-Agent Policy Gradient (COMA, Foerster et. al. 2017). Ultimately, the counterfactual baseline for the advantage function provided by COMA was the primary update incorporated into the PPO algorithm, also to the greatest effect. Teamwork, and success in the multi-agent environment, was measured in concepts like the tiredness in the players and the number of passes, expected to decrease and increase respectively, along with more general concepts like win ratio. The changes made to PPO were to mixed effect, but certainly this is subject matter I would be ecstatic to return to in pursuit of greater results.