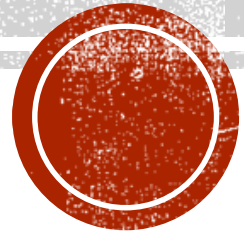
RESULTS FROM EXPERIMENTS

Skanda Vaidyanath

Institute of Creative Technologies

6/19/2019

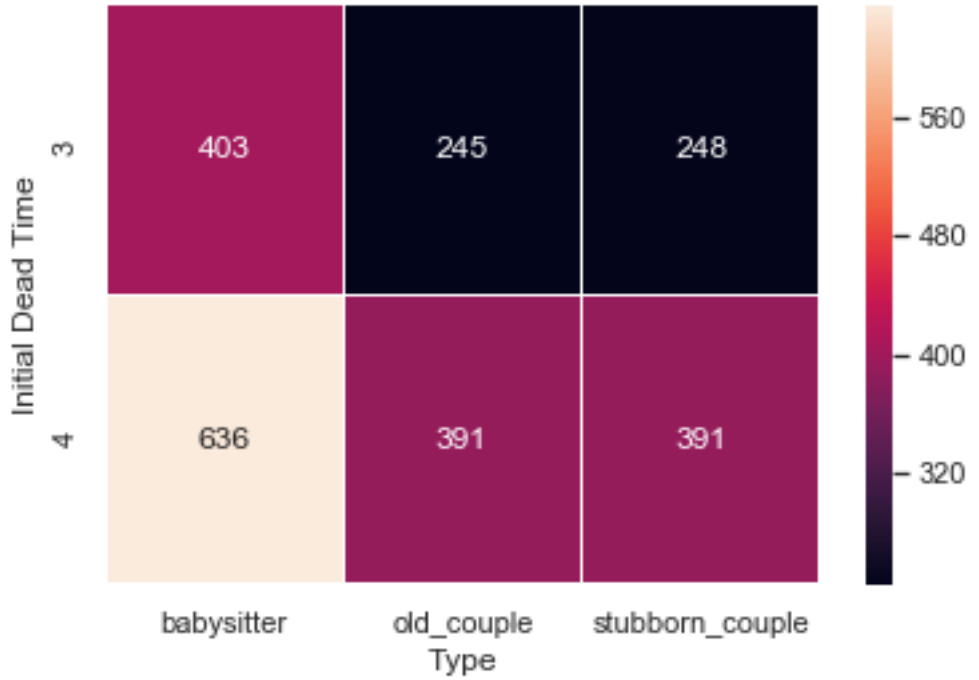


FAVORABLE INITIAL CONDITIONS

- The dead time can only be initialized with a value of 3 or 4
- Algorithms are Q-learning, SARSA, Expected SARSA, Monte Carlo, Off Policy Monte Carlo.

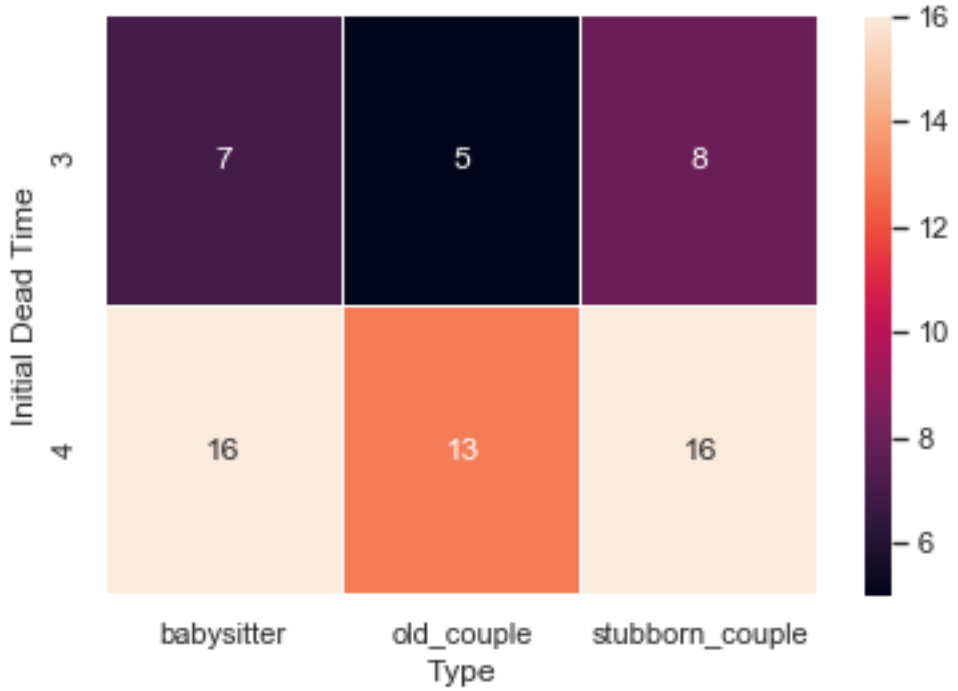
Q-LEARNING

- Training statistics



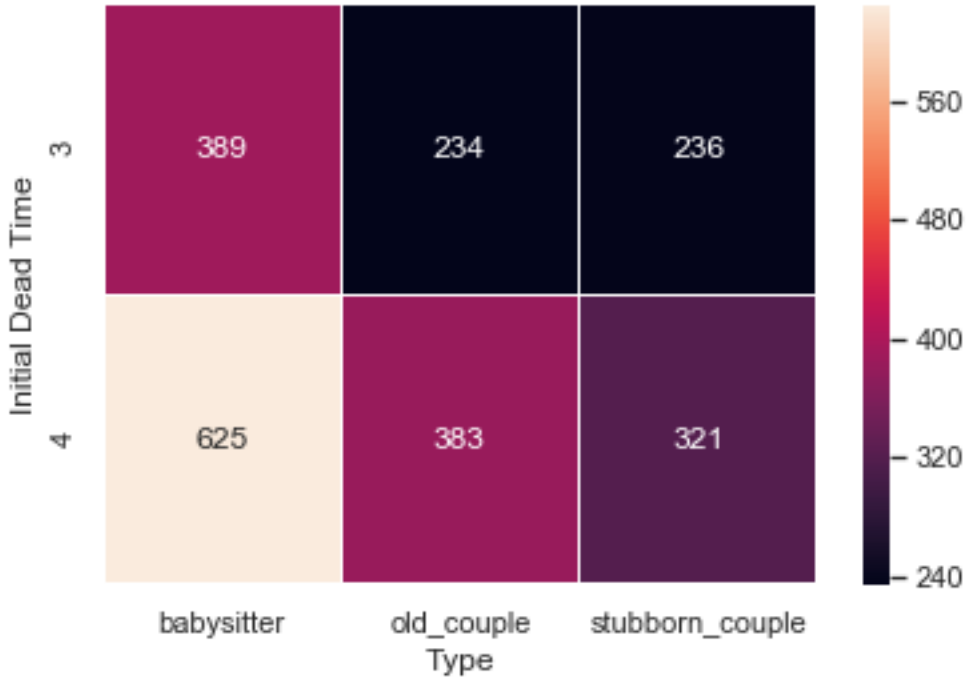
Q-LEARNING

- Testing Statistics



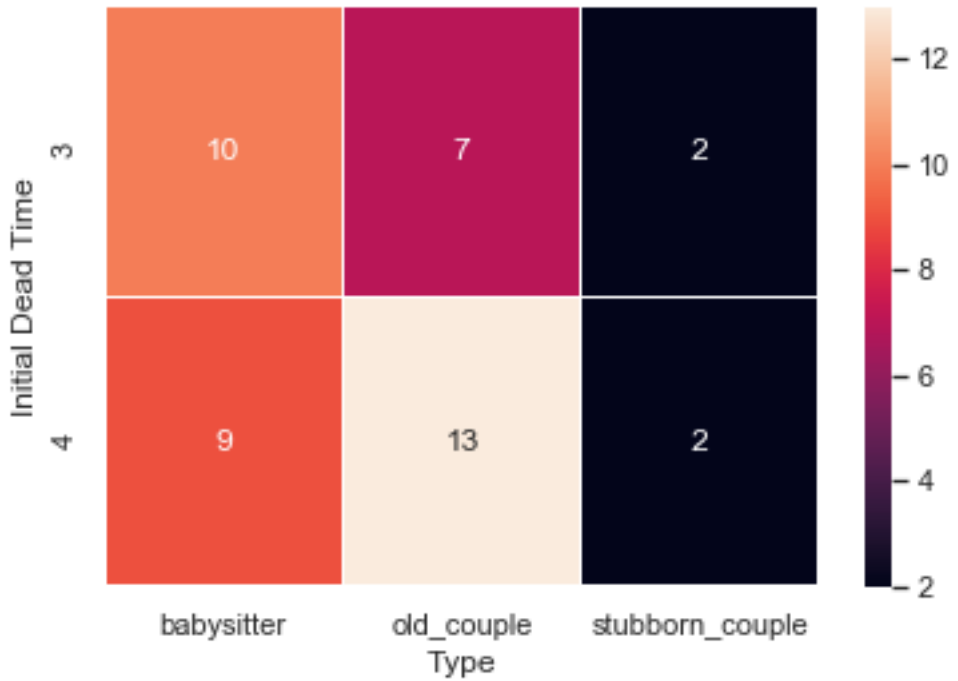
SARSA

- Training statistics



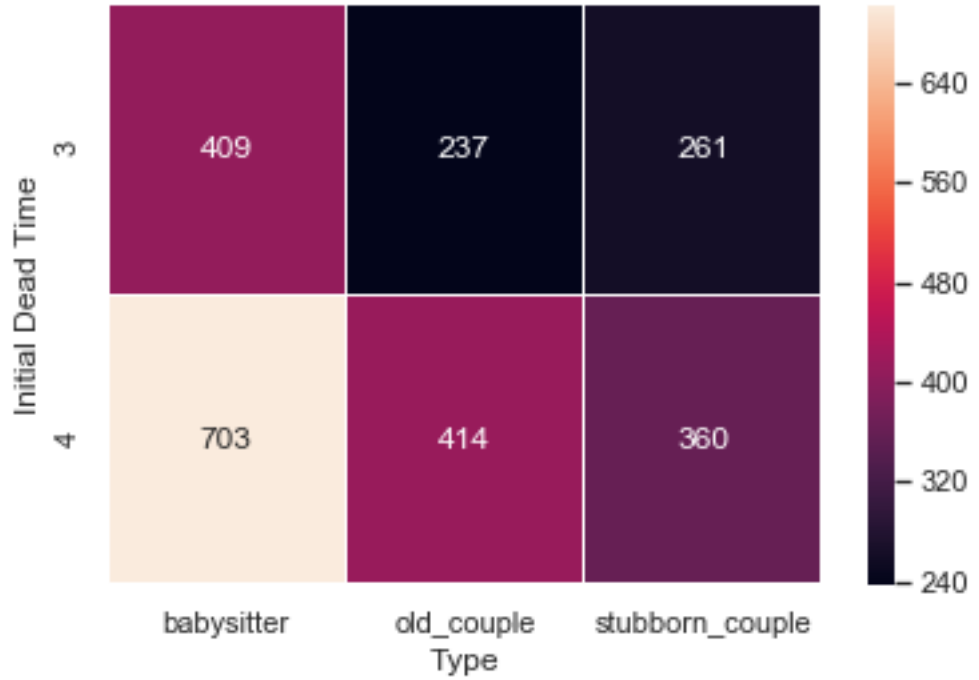
SARSA

- Testing Statistics



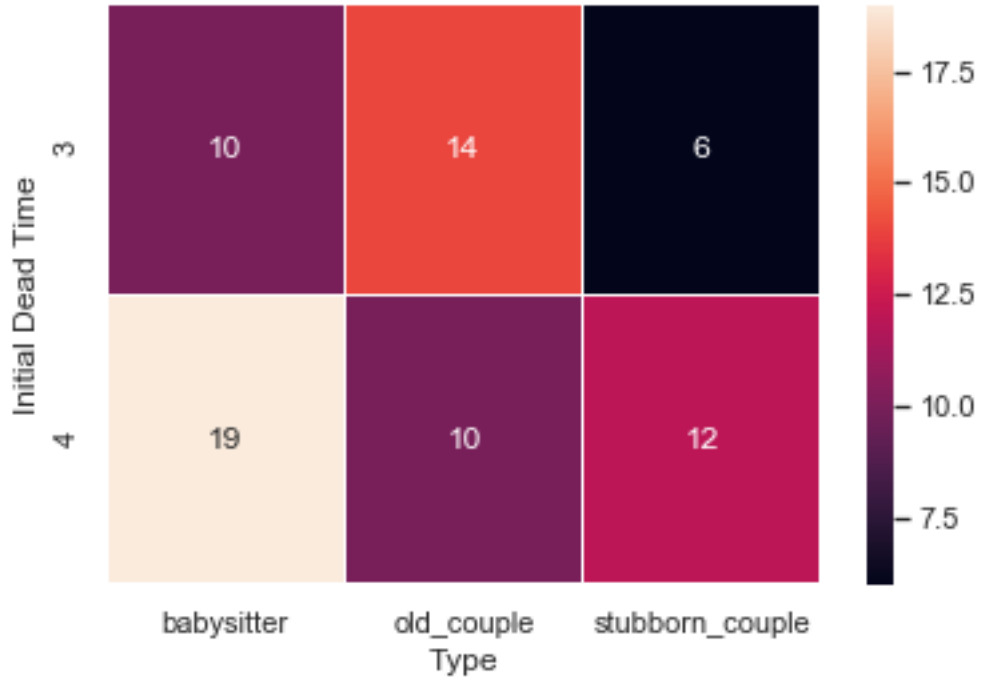
EXPECTED SARSA

- Training Statistics



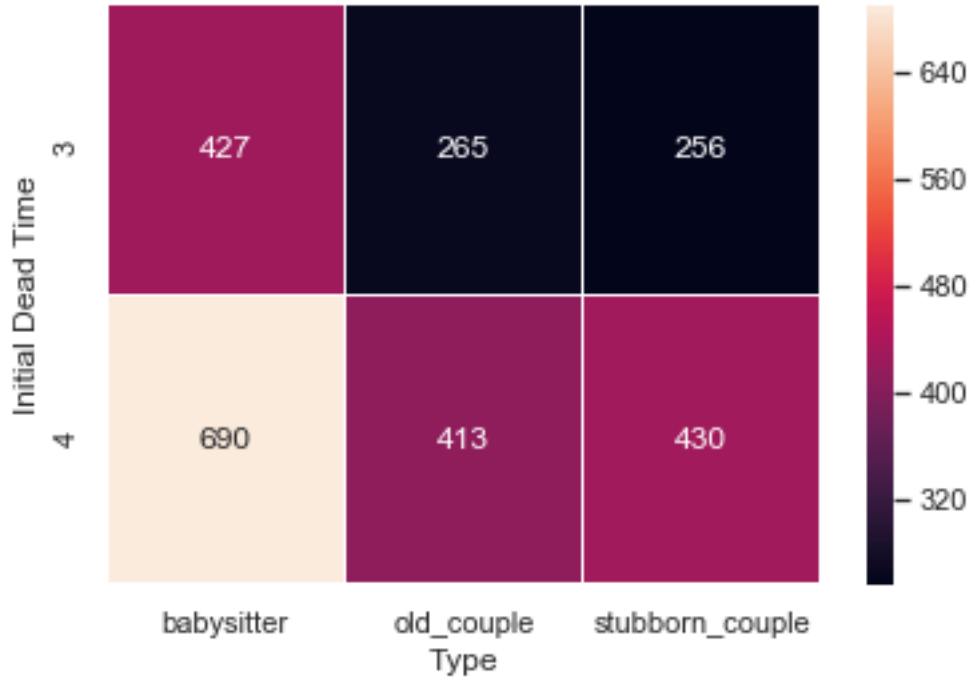
EXPECTED SARSA

- Testing Statistics



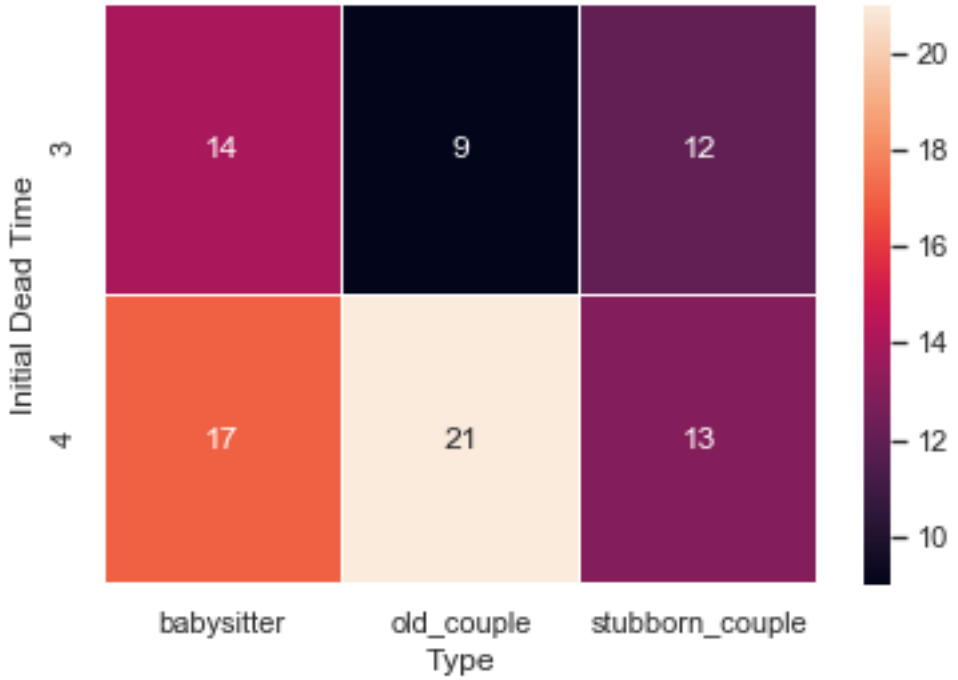
MONTE CARLO

- Training statistics



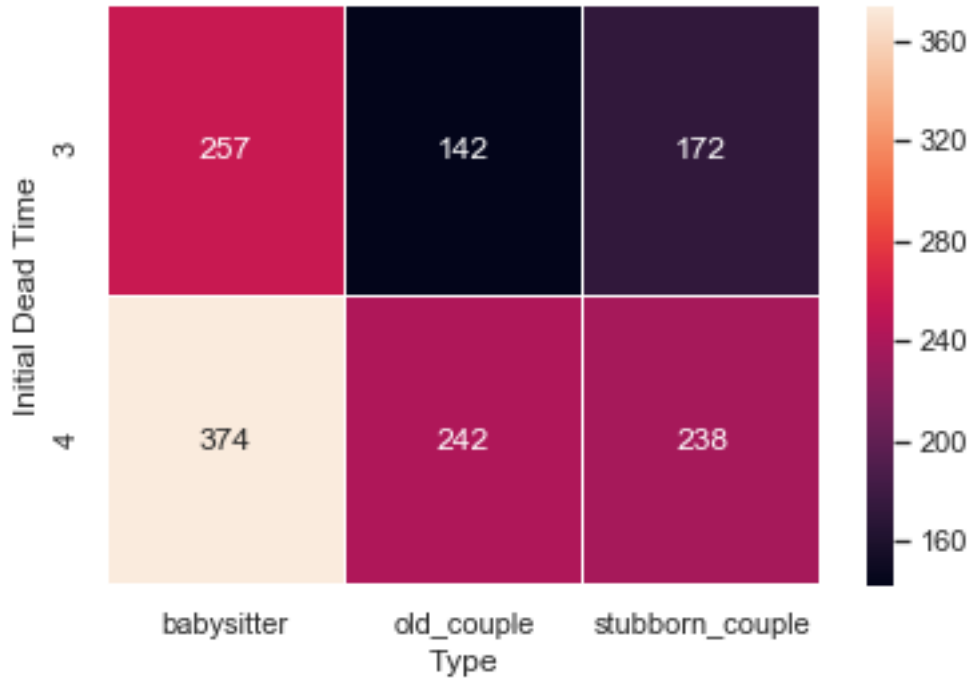
MONTE CARLO

- Testing statistics



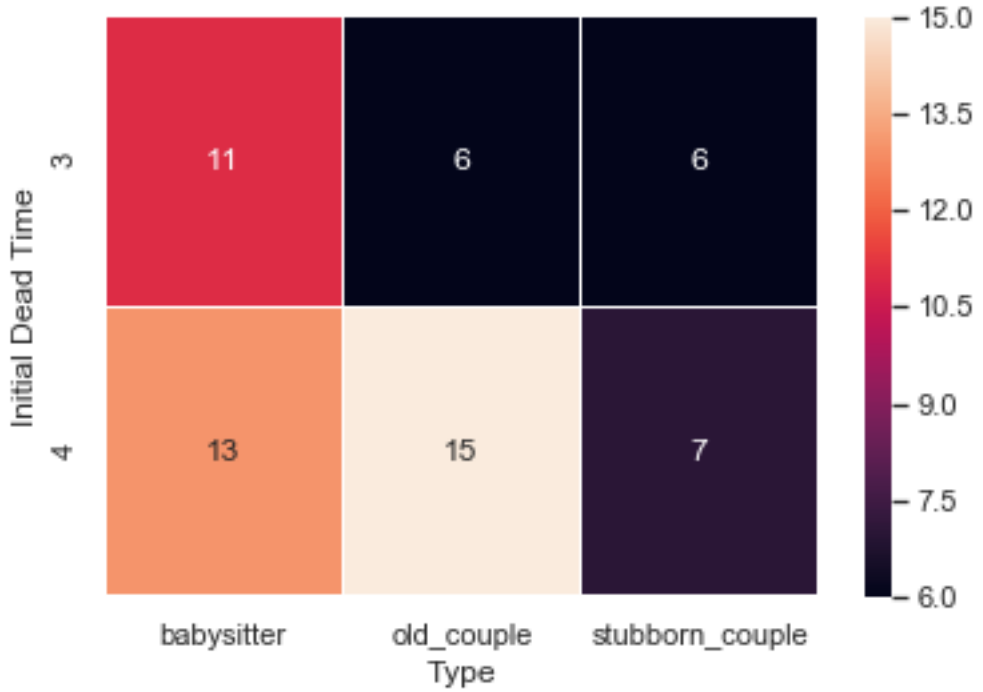
OFF POLICY MONTE CARLO

- Training Statistics



OFF POLICY MONTE CARLO

- Testing Statistics



OBSERVATIONS

- All models are able to save more people in this setting
- All the models have learned to interrupt the operator all the time, either in the very beginning or by terminating an on-going negotiation by interrupting the operator. With the current reward function in place, this is not a bad thing to learn because the operator can save any of the three civilian groups with a 100% guarantee. If we want to stop this from happening, we need to augment our reward function.
- The models have learned when to query for guide and when to call the appropriate guide. They still play them at "bad times" but again, to stop this, we need to augment the reward function
- In this version, models seem to be terminating a warning midway for a SPN or a SPN midway for a warning or an ON etc.
- The models still make some weird moves like WAIT as the very first action or GUIDE as the very first action, etc.
- ALL MODELS HAVE ONLY BEEN TRAINED FOR 10,000 EPISODES. For such a sparse reward function, I think we may need to train them for more episodes.