1. Description
   1. Using strategies borrowed from dynamic programming, construct computational models that will respond to cues presented by the environment in the same way as actual participants. Then, using these models, attempt to determine the contributions of various decision making algorithms, and assess their correlation with reaction times.
2. Background on reinforcement learning
   1. Original motivation came from computer scientists in the 1950s, who were trying to teach computers to play chess using what they termed an evaluative function. Bellman conducted some of the first research in the 1950s
   2. In computer science, this family as algorithms known as dynamic programming or what they termed in the 1950s establishing “optimal control”. In cognitive neuroscience, it’s known as reinforcement learning. IT doesn’t matter what you call it – they both attempt to address same problem – how to create an autonomous system that will respond optimally to maximize reward over a long period of time
   3. A definition – the agent is the entity who is learning
   4. Does this by mapping discrete situations to actions to take in those situations. This is mapping known as the policy. The agent simultaneously tries to maximize a numerical reward signal
   5. Characterized by
      1. Trial and error search – agent is not told which actions to take, but instead must balance between exploitation and exploration
      2. NOT supervised learning. No teaching signal – agent must explore on its own
   6. Four main elements
      1. Policy – mapping from states to actions previously discussed
      2. Reward Function – maps each state to a reward value. Is short sighted
      3. Value Function – corrects short sightedness of the reward function by specifying what’s good in the long run. Value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state.
   7. This next and final element is crucial to my project. Some agents also contain a :
      1. Model of Environment (Optional) – internal representation that mimics the behavior of the environment. Given a state and an action, model might predict the resultant state and the next reward. Models are the basis of planning. Model based reinforcement learning is a relatively new development
      2. Agents that include a model are known as model based
      3. Agents that do not include a model are known as model free
      4. These two scenarios are the various learning algorithms I described above, and we’ll revisit them
      5. State action diagram – consider a diagram with more choices shown than just the optimal one. Imagine trying to walk down all the possibilities in a manner analogous to a breadth first search. Very time intensive
      6. Balloon example
         1. Model based – has an internal model of the invironment, which he can consult and manipulate to determine the best possible action. Imagine the different possibilities – moving to a different room, muffling the pop, and so on. Must compute best possible move
         2. Model free – rewards for actions are already stored. Similar to instinct. All he has to do is access the cached values
3. My Project – revisited from the start of the presentation
   1. Data provided by Nathaniel Daw, Neuroscience professor at NYU. Contains seventeen different subjects response times and decision patterns over two hundred different trials
   2. Build agents that can make decisions in the same way as the actual participants
   3. Determine degree to which subject’s responses were determined by model-free and model-based decision processes.
   4. Determine the correlation between proportion of model free/model based learning and reaction times
4. Motivation – why look at reaction time?
5. Reaction time seems like the natural choice
   1. Can be considered in computer science terms
   2. Model free agents should be able to respond quickly. The values are already pre-cached – all the agent has to do is access them. It’s analogous to the quick read times when loading from RAM
   3. Model based learning requires more computational muscle, and therefore should take longer. Consider the breadth first search previously discussed
   4. Supported by recent study done at Princeton – people become more model-free when conducted under computational load
6. Why?
   1. Important implications for how humans learn/ make decisions
   2. Better computational models could point towards important empirical experiments
   3. Convergence of fields – drift diffusion models predict different reaction time distributions for different learning models,
   4. Simply has not been done – very new field
7. Daw’s study
   1. On each trial, an initial choice between two options labeled by (semantically irrelevant) Tibetan characters led probabilistically to either of two, second-stage ‘‘states,’’ represented by different colors. In turn, these both demanded another two-option choice, each of which was associated with a different chance of delivering a monetary reward. The choice of one first-stage option led predominantly (70% of the time) to an associated one of the two second-stage states, and this relationship was fixed throughout the experiment. However, to incentivize subjects to continue learning throughout the task, the chances of payoff associated with the four second-stage options were changed slowly and independently, ac-cording to Gaussian random walks. Theory ( Daw et al., 2005; Dickinson, 1985) predicts that such change should tend to favor the ongoing contribution of model-based evaluation.
   2. The logic of the task was that model-based and model-free strategies for RL predict different patterns by which reward obtained in the second stage should impact first-stage choices on subsequent trials. For illustration, consider a trial in which a first-stage choice, uncharacteristically, led to the second-stage state with which it is not usually associated, and in which the choice then made at the second stage was rewarded. The principle of reinforcement would predict that this experience should increase the probability of repeating the first-stage choice because it was ultimately rewarded. However, a subject choosing instead using an internal model of the task’s transition structure that evaluates actions prospectively would be expected instead to exhibit a decreased tendency to choose that same option. This is because any increase in the value of the rewarded second-stage option will more greatly increase the expected value of the first-stage option that is more likely to lead there. This is actually the first-stage option that was not originally chosen.
8. Why?
   1. Important implications for how humans learn and make decisions
   2. Better computational models could point towards important empirical experiments
   3. Convergence of fields – drift diffusion models predict different reaction time signatures for different learning models
   4. Simply has not been done – very new field
9. Further Research
   1. Use reaction times as constraints to build more accurate computational models
   2. Analyze distribution of reaction times (reaction time signature) to determine which of many model-based learning algorithms the subject used, combining with insights from drift diffusion models