Learning in uncertain environments: comparing human and macaque reinforcement learning against an optimal benchmark and quantitatively measuring reward prediction error in BOLD.

*E. J. DeWitt[1], M. Dean[2], P. W. Glimcher[3];

[1]Psychology, New York Univ, New York, NY, [2]Economics, New York Univ, New York, NY, [3]Center for Neural Science, New York Univ, New York, NY.

How we learn the expected value of an action is a fundamental question for neuroscience, microeconomics and computer science, but one that has been difficult to study in a normative context. Prior behavioral research in restricted dynamic learning environments (e.g. repeated games, dynamic foraging tasks) has not been able to leverage the extensive theoretical, behavioral—and recent neurobiological—research on reinforcement learning. Prior neuro-psychological studies of reinforcement learning in humans have used tasks that precluded comparison to an known optimal behavior. We have developed a task that can present a wide range of dynamic learning environments with a known optimal policy for learning identical to classic reinforcement learning models. This task allows us to relate observed human choice behavior during reinforcement learning both to underlying neural mechanisms and to normative theory at the same time.

We used our task to compare human and macaque choice behavior to that of an ideal decision maker in a variety of environments that call for different optimal learning rates. Our task, an n-armed drifting bandit, allows us to select task parameters that specify any optimal learning rate we desire. In the task, subjects face two or more 'arms' that each return random rewards, the magnitude of which (the *reward magnitude*) is drawn from Gaussian distribution. When an arm is sampled, the mean of this distribution drifts unpredictably (the *drift rate*, also drawn from a Gaussian). The functional form of the recursive optimal solution to such a bandit is identical to Q-Learning, with the optimal learning rate specifiable using a Kalman filter. The Kalman gain is thus used to determine the learning rate analytically based on these two parameters. Importantly, these two parameters trade-off. Increasing drift rate and increased variance can cancel to yield a constant optimal learning rate. A Gittins' index based on this analytically determined optimal learning rate (the optimal Kalman gain) can then be used to specify the best action on a choice-by-choice basis.

Do humans and animals adjust their learning rate appropriately when the optimal rate changes? We found that while humans employ a learning rate that is too high (recent reinforcements are over-weighted), they do so in a consistent manner. When changes in

the environment call for higher or lower learning rates, humans do significantly increase or decrease their learning rates. Heavily overtrained rhesus macaques, however, are almost optimal in adjusting their learning rates. Interestingly, human efficiency can be improved by loading working memory (with a concurrent n-back memory task) while humans perform the bandit task. Under these conditions human performance more closely approximates that of the monkeys. It is not known if overtraining the humans would have a similar effect. Because we found that human's are, in the absence of overtraining and a working memory load, inefficient in their reinforcement learning, we are led to ask the question: Does the activity of the mid-brain dopamine nuclei, thought to underlie reinforcement learning in animals, reflect this inefficiency or are dopaminergic circuits efficient but overruled by other structures? Prior work in animals and humans has shown that dopamine activity correlates with reward prediction errors. However, to measure within subject learning efficiency and compare it to dopamine activity, we require an approach that allows us to measure the influence of rewards on dopamine activity directly. Bayer et al 2005 demonstrated that dopamine activity measured using single unit electrophysiology could be shown to quantitatively be calculating a reward predication error. Using similar linear methods we have demonstrated within subject behavioral changes in learning using a variant of our reinforcement learning task designed to address the requirements of quantitative BOLD measurements using fMRI. We are using these methods to compare behavior and neural reinforcement learning within individual human subjects.

Funding Contributed by: JSMF21002079