Motivation: How do we update our expectations of the world in response to new information? Canonical theories of learning assume we integrate information in an unbiased fashion, weighing all feedback equally. Why is it then that we are sometimes stubbornly resistant to criticism, but other times receptive only to it? Could it be that prior expectations actually change how we learn, biasing our sensitivity to feedback in favor of information corroborating or contradicting our expectations? In my research, I propose to utilize behavioral experiments, computational models, and fMRI to investigate (1) how expectations bias new learning; (2) when we learn more from feedback consistent with prior expectations (confirmatory bias) vs. from feedback contradictory to expectations (surprise bias); and (3) how these processes are instantiated in the brain.

The computational framework of reinforcement learning (RL) offers a precise means of measuring and quantifying the dynamics of trial and error learning.[1] In RL, the discrepancy between predicted and experienced outcomes (e.g. expecting criticism, but receiving praise) is quantified as the reward prediction error (RPE). Future expectations about outcomes are incrementally learned by multiplying the RPE by a “learning rate” parameter, which controls the extent to which a single outcome affects future expectations. Importantly, biases in learning can be measured as asymmetric learning rates, wherein a learner exhibits greater learning from positive RPEs (unexpected praise) or from negative RPEs (unexpected criticism). Evidence for asymmetric learning in humans is mixed, with studies reporting both positive and negative RPE biases.[2-5] However, these studies did not control for the potential impact of prior expectations on learning rates. In the proposed research I will explicitly manipulate prior expectations of rewarding and aversive outcomes during a learning task and evaluate, using RL models, if this causes asymmetric learning from positive and negative feedback. I will also assess the direction of the asymmetry to identify if learners discount or upweight prediction errors contradictory to their expectations. I hypothesize that (H1) prior expectations will bias learning as measured by asymmetric learning rates; and (H2) human learners will exhibit individual differences in the direction of this bias, as predicted by stable trait differences and neural activity.

Another advantage of using the RL framework is that it has been robustly mapped to the brain regions likely involved in these learning processes. Dopaminergic neurons in the brainstem encode RPEs and project these to the striatum,[1] which in turn exhibits asymmetric activity during biased feedback learning.[2] Cortical regions like the ventromedial prefrontal cortex (vmPFC) and insula are thought to encode expectations about rewards and punishments, respectively, and signal this information to the striatum.[6] Noradrenergic neurons in the locus coeruleus (LC) are implicated in modulating learning rates in response to contradictory feedback[7] and in promoting new information over prior expectations in evidence accumulation tasks.[8] Given these structures exhibit strong anatomical connections[7], it is plausible their combined activity shapes learning in the striatum. Using fMRI, I will test the hypotheses that (H3a) asymmetric RPEs in the striatum are consonant with individual learning biases, and (H3b) the direction of the asymmetry reflects patterns of functional connectivity between the striatum, cortex (vmPFC, insula), and LC, where cortex biases learning in favor of prior expectations and LC biases learning in favor of surprise.

Methods & Analysis: To measure the role of prior expectations in biased feedback learning, I designed a probabilistic learning and decision-making task involving both monetary rewards and losses. On every trial, participants choose between two available options (of four total). The outcomes associated with each option are drawn from independent Bernoulli distributions with nonstationary (drifting) means, such that the likelihood of experiencing gain (+\$0.25) or loss (-\$0.25) from any option varies over time, requiring constant learning of which option is best.

Importantly, the drifts are predetermined so as to create periods of overall high average reward and periods of high average punishment. I expect participants will form corresponding positive and negative expectations, which I will confirm by assessing periodic self-report estimates of the average rate of reward. If asymmetric biases in feedback learning reflect prior expectations, then I expect the asymmetry to reverse in direction during shifts from high average reward to high average punishment (and vice-versa).

In Experiment 1, I will test hypotheses H1 and H2. To obtain the statistical power needed to characterize individual differences in learning (H2), I will recruit a sample of 500 participants from Amazon Mechanical Turk. I will fit Bayesian hierarchical RL models to participants' choice behavior to quantify the relationship between expectations and asymmetric learning. To reiterate, I predict that individuals' learning rates for positive and negative RPEs will differ in magnitude and that these asymmetries will reverse across experimental contexts (high reward vs. punishment). I also predict that I will find stable heterogeneity in the direction of these biases across individuals. For example, I expect some participants will exhibit greater learning from positive RPEs in the reward context (confirmatory bias) whereas others will exhibit greater learning from positive RPEs in the punishment context (surprise bias).

In Experiment 2, I will test hypothesis H3 (and replicate H1 and H2). In a second cohort of 60 participants, I will collect simultaneous functional magnetic resonance imaging (fMRI) and pupillometry as they perform the same task. fMRI will be used to measure activity in the brain regions described above during choice and learning. Pupillometry will serve as a secondary measure of LC activity.[7] I will fit the same computational model to the choice behavior of these participants and use variables from the model to perform model-based fMRI analysis. To measure functional connectivity, I will also perform psychophysiological interactions analysis. I predict the striatum will exhibit asymmetric RPEs in biased learning, and that functional connectivity between the vmPFC/insula/LC will predict the direction of the bias.

Intellectual merit: Virtually all learning problems that humans must solve involve prior expectations. The proposed research will provide novel insights into how these expectations shape learning in ways that may confirm or contradict those very expectations, with wide implications to learning in cognitive or social domains. Our results will resolve inconsistencies in the RL literature by demonstrating a link between reward expectations and learning asymmetries. Finally, our findings will also illuminate the neural mechanisms underlying these processes, bringing ideas from sensory evidence accumulation to bear on learning and decision making.

Broader impact: Biased feedback learning is implicated in psychiatric illness including anxiety and depression.[4] Determining how expectations shape new learning has tremendous value for cognitive behavioral therapy (CBT), the leading treatment for anxiety and depression, in which patients address the pessimistic expectations central to their illness. My research may lead to improved methods for CBT, such as utilizing surprise during treatment to counteract learning biases and prescribing norepinephrine-reuptake inhibitors to promote new learning.

Feasibility and support: My advisors, Drs. Yael Niv and Nathaniel Daw, are leaders in the field of reinforcement learning. Both have extensive experience in computational modeling, fMRI data analysis, and large-scale data collection, as well as in collecting and analyzing pupillometry data. In addition, I am well-versed in computational modeling and fMRI analysis (see personal statement). Princeton has 2 research-dedicated 3T MRI scanners with integrated pupillometry.

References: 1. Niv (2009, *J Math Psychol*); 2. Niv et al. (2012, *J Neurosci)*. 3. Gershman (2015, *Psychon Bull Rev*). 4. Sharot & Garrett (2016, *TICS*); 5. Lefebvre et al. (2017, *Nat Hum Beh*). 6. Palminteri et al. (2015, *Nat Comm*) 7. Sara et al. (2009, *Nat Rev Neurosci*); 8. de Gee et al., (2017, *eLife*).