**Methods**

*Participants*

A total of 703 subjects were recruited on Amazon Mechanical Turk to participate in Markov decision tasks. Each subject participated in only one task. Subjects were excluded from analysis if they timed out on more than 50 trials, or if their final accumulated score was below zero. After applying our exclusionary criteria, there were 217 subjects and 6090 critical trials in Experiment 1a, 243 subjects and 6206 critical trials in Experiment 1b, 49 subjects and 3089 critical trials in Experiment 2a, and 194 subjects and 3169 critical trials in Experiment 2b.

*Experiment Design*

The designs of Experiments 1-3 are depicted in Figs. 1A, 3, and 4A. The two Stage 1 options for each trial were always chosen such that the options led to different Stage 2 states (i.e. (1,3) were never paired in Experiment 1). All rewards distributions were initialized uniformly at random on a range of -4 points to +5 points, and varied according to a bounded Gaussian random walk for the remainder of the experiment. After each round, the drift was sampled from a normal distribution with (μ=0, σ=2), rounded to the nearest integer, and added to the current reward level. In cases where drift selected a reward level outside the bounds of [-4,5], the reward would ‘rebound’ by the amount of the excess. The rewards on setup trials (those immediately preceding critical trials) were boosted to their extremes by adding +2 or -2 points, depending on the reward distribution’s current sign. If the boost selected a reward level outside the bounds, the reward remained at the boundary amount.

After the experiment, participants received a bonus payment based on their accumulated points. Each point was worth 1 cent. Participants were informed of the value of points in the instructions. Each participant completed 75 practice trials followed by 175 rewarded trials. The practice trials were divided into three sections of 25 practice trials each. Sections were designed to ease participants into the task by introducing one task element at a time. On the rewarded trials, subjects had only 4 seconds to make their choice between the two numbers. If they did not make a choice within 4s the trial would time out and the next trial would begin. Practice trials had no time limit. Participants in Experiments 1a and 1b saw 26 critical trials each. The spacing of critical trials in Experiments 1a and 1b was chosen randomly, with the constraint that they had to be at least three trials apart from each other.