**Task:** Participants first complete 520 trials of a forced choice learning phase to ensure they know the state space and transitions well, and that each participant sample the same number of times from each state transition in the state space. Every 30 or so trials during this forced choice learning phase, participants are given a memory quiz where they must choose one of the two first-stage states that has the best chance of transitioning them to a 3rd stage state. We ensure in the piloting that our state space produced very good memory of the state space (average XX% on these memory probes).

After this planning phase, participants face planning queries that ask them to use what they learned about the probability of transition to certain states in the preceding training phase to win the most money. We explicitly tell participants that they should consider going for smaller rewards if they believe they have a much greater chance to reach those states than more highly rewarded but rare states (i.e., we encourage them to incorporate both the instructed reward magnitude and learned probabilities of state transitions into their planning process).

For each planning query, 1 or 3 reward locations are instructed on the screen, and the participant is told to use this reward information to choose an action that will maximize their reward. In order to do so, they should use what they have learned about the state space in the preceding training phase, which determines the likelihood of transitioning to successive states given an initial action. Participants know that if they reach a location, they will receive the reward found there (i.e., the reward function is deterministic). However, participants are never shown whether they receive the reward; they are merely told the computer is simulating whether they in fact reached the reward, and this will affect the bonus money they receive at the experiment’s conclusion. This is intended to keep learning unaffected by reward receipt, removing the possibility that model-free learning impacts subsequent decisions. Participants are given unlimited time to deliberate before taking an action.

To maximize our chances of supporting our main hypothesis that participants utilize PR-based planning over SR-based planning, we designed the state space to include initial actions that differ in base-rates, which only PR-based planning is sensitive to. Thus, the low base-rate starting state was experienced 160 times whereas the high base-rate starting state was experienced 360 times. The degree to which the base-rates differed was determined in simulation to allow for planning queries by which SR preferred low base-rate actions and PR-based planning preferred high base-rate actions.

To test our main hypothesis, we designed 4 single-goal planning queries wherein PR-based and SR-based planning arrive at different conclusions regarding which action is best to take, but holding constant across strategies the ratio of best:worst action value (detailed in Table 1). We opted for single goal queries because it removes the possibility by which subjects only utilize a subset of reward or probability information presented in multi-reward planning queries, which could make it more difficult to easily compare SR- and PR-based planning, and which we observed in our previous Sudy 1 (OSF CITE) data when subjects indeed faced multi-reward planning queries. More generally, the use of single-goal queries allows us to account for 24 possible instantiations of PR- and SR-based planning, including agents that consider only the highest state bearing the largest reward or probability magnitude when planning, and using a variety of ways to then select the best action to reach that state (including comparing expected values, reward magnitude, absolute or binarized PR- and SR-based probabilities; see Appendix for full list of algorithms in the model set).

Given our previous demonstration in Study 1 that a diverging state space incentivizes the use of PR-based planning, we retained a diverging state space in the present study, which meant that it was not possible to equate the magnitude of action value differences on single-goal queries. However, we ensured single-goal queries incentivized the SR- or PR-based optimal action to the same degree by equating the *ratios* of expected values according to SR- and PR-based planning. As an example, whereas a planning computation based on using a PR representation of the state space might predict action 1 is most likely to acquire the most reward by a reward ratio of 4:2, planning based on an SR representation would predict action 2 is most likely to acquire the most reward by an equivalent ratio of 2:1.

As in Study 1, these predictions of action value differences according to PR-based planning and SR-based planning were carried out via simulation studies. In our simulations, PR and SR matrices were derived by an agent that sampled transitions under a random policy, and computed the observed transition frequencies via recursive updates with a learning rate that declines with each sample from a given state-action pair. Because this sampling was done extensively (over 10,000,000 experiences), the resultant PR and SR matrices very closely approximate the true PR and SR probabilities.

In addition to our primary hypothesis that PR-based planning would be used instead of SR-based planning, we additionally sought to replicate our exploratory finding in Study 1 that individuals select only the highest reward as a goal to plan for using PR-based probabilities (as opposed to SR-based queries, which was the second-best fitting model; difference in average log-likelihood = 30), and thus neglect information about other rewards on offer. Thus, to test whether participants use PR-based probabilities to plan for the highest goal, we crafted 4 specific 3-reward queries (see Table 1), in which a PR-based strategy which favors the highest reward predicts one action, and a PR-based agent that integrates information over all rewards on offer (by averaging over their PR-probability\*reward magnitude products) favors the alternate action.

To rule out that performance on the queries designed to test our main hypothesis could be explained by bias to select the action with the higher base-rate, we designed 4 single-goal queries that according to either PR- or SR-based planning on average preferred neither action. Specifically, PR- and SR-based planning predict the low base-rate action on half the queries, thus resulting in the prediction that on average, PR- and SR-based planning prefers neither action.

Note, we finally will rule out MB-based planning as was done in Study 1 with a similar transition revaluation manipulation, in which 2nd-3rd state transitions are altered via instructed, followed by subsequent planning that can only be carried out by an agent using MB-based planning.