Humans rationally balance detailed and temporally abstract world models

Ari E. Kahn[1] and Nathaniel D. Daw[1,2]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[2]Department of Psychology, Princeton University, Princeton, NJ, USA

July 9, 2024

Abstract

How do people model the world's dynamics to guide mental simulation and evaluate choices? One prominent approach, the Successor Representation (SR), takes advantage of temporal abstraction of future states: by aggregating trajectory predictions over multiple timesteps, the brain can avoid the costs of iterative, multi-step mental simulation. Human behavior broadly shows signatures of such temporal abstraction, but finer-grained characterization of individuals' strategies and their dynamic adjustment remains an open question. We developed a novel task to measure SR usage during dynamic, trial-by-trial learning. Using this approach, we find that participants exhibit a mix of SR and model-based learning strategies that varies across individuals. Further, by dynamically manipulating the task contingencies within-subject to favor or disfavor temporal abstraction, we observe evidence of resource-rational reliance on the SR, which decreases when future states are less predictable. Our work adds to a growing body of research showing that the brain arbitrates between approximate decision strategies. The current study extends these ideas from simple habits into usage of more sophisticated approximate predictive models, and demonstrates that individuals dynamically adapt these in response to the predictability of their environment.

Introduction

A hallmark of human and animal planning is its flexibility: the ability to plan and replan effectively in the face of novel or changing circumstances and outcomes[1]. This requires not only predicting the immediate outcomes of actions based on experience, but also inferring longer-term payoffs, which often depend on a series of subsequent states and actions[2]. A long-considered solution to this problem is for the brain to simulate outcomes through use a cognitive map or internal model[3]. Reinforcement learning (RL) theories formalize this idea in terms of "model-based" (MB) algorithms[4], which employ a learned model of the short-term consequences of actions in order iteratively to simulate the long-run consequences of candidate actions. Other learning approaches aim to simplify this laborious decision-time computation at the expense of reduced flexibility. Notably, model-free (MF) RL algorithms[5] directly learn and cache the long-run aggregate reward expected for performing an action (the key decision variable), without representing or computing over the step-by-step contingencies. This avoids costly model-based search, but at the expense of sub-optimal behavior in some circumstances: for instance, when distal contingencies or goals change in a way that invalidates the cached aggregates. It has been suggested that the brain employs both methods, and that it judiciously trades them off, balancing their costs and benefits according to situational demands[6–9]. This approach suggests a formal, resource-rational account of both healthy[10] and pathological habits[11–13], as well as helping to characterize changes in their deployment in other situations such as over development and aging[14,15].

The successes of this framework as well as the practical limitations of both extreme algorithms[16] have motivated the study of more nimble methods that strike an intermediate balance between flexibility and computational cost. One important class of approaches simplifies internal models and model-based prediction by leveraging *temporal abstraction* over future states[17]. Prominent among these is the Successor Representation (SR)[18], a planning algorithm that models future states (instead of future rewards) aggregated over multiple timesteps. Like MB RL with a one-step model, such a multi-step world model allows computing or recomputing net payoffs (by combining state predictions with up-to-date per-state rewards) — but without the computational expense of

step-by-step simulation[19]. However, the modeling assumptions that simplify prediction (forecasting multiple steps ahead at once) are not always valid: if intermediate contingencies or preferences change in a way that invalidates the SR's summarized world model, it will misvalue actions[20]. Such characteristic slips of action both impose a potential cost on SR usage (offsetting gains due to streamlined computation) and render its use empirically detectable. Indeed, a number of studies of human and animal decision making suggest that the brain uses this or similar temporally abstract predictive representations, at least some of the time[20–27]. However, it remains unclear whether the concept of resource-rational trade-offs[28] (e.g. balancing the costs and benefits of SR vs. full MB as appropriate to the circumstances) can be extended to this case. Furthermore, this is in part an empirical issue, in that the clearest evidence supporting SR usage comes from unwieldy one-shot task designs that use a single-choice probe (following laborious training and retraining phases) to elicit slips of action that distinguish SR from full MB[20]. This makes it difficult to measure between-individual or -condition variation in SR usage, as well as tracking other more dynamic aspects of trial-by-trial SR learning and associated neural signals.

To study within-subject dynamics of SR usage, one promising approach is to examine its more granular, dynamic predictions about how past actions and outcomes guide future behavior on a trial-by-trial basis. This strategy has proven helpful in dissecting MB versus MF behavior, using two-step Markov decision tasks that expose differences in how MB vs. MF approaches update their choice preferences, trial by trial, in light of each trial's outcome[29]. Here, we pursued a parallel approach by designing a novel multi-step decision task that elicits unique trial-by-trial behavioral signatures of each planning strategy, reflecting recursive maximization for MB and on-policy state occupancy averaging for SR. Leveraging these signatures, we can characterize individual subjects' relative reliance on either approach, as a weighted mixture of the two strategies: estimating value via the multi-step state expectations derived from prior experience (SR), or a full tree search that maximizes at each step (MB).

The ability to characterize graded SR vs. MB usage within individuals allows us next to investigate whether people rationally adjust their reliance on the SR in response to circumstances such as environmental predictability. Although a line of theoretical and experimental work has focused on the judicious deployment of MB vs. MF[6,7], a recent line of theory suggests that a similar (and potentially more fundamental) tradeoff may play out in the dynamic tuning of world-model abstraction between SR and MB[30,31]. Whereas the utility of MF learning versus MB planning depends on the stability of action value estimates[32], the predictive utility of the SR should analogously depend on the stability of long-run future state occupancy. We hypothesized that individuals would dynamically adjust their usage of the SR to the local temporal statistics of the task, and in particular that choices would increasingly reflect the SR when its state predictions were more stable and reliable. We estimated the balance of MB and SR contributions to behavior while manipulating these temporal statistics, and found that people rationally adjusted usage to match the accuracy of the SR. This suggests that humans can arbitrate between a temporally abstract model and full MB planning in response to their respective utility, adding to and recentering a growing literature on rational balance between decision mechanisms.

Results

Task

Participants (n=100) completed a two-stage Markov decision task in which they repeatedly chose between two islands and then between one of two pairs of boats, each boat with an unsignaled probability of reward. Their goal was to maximize their total monetary payout by selecting the boats

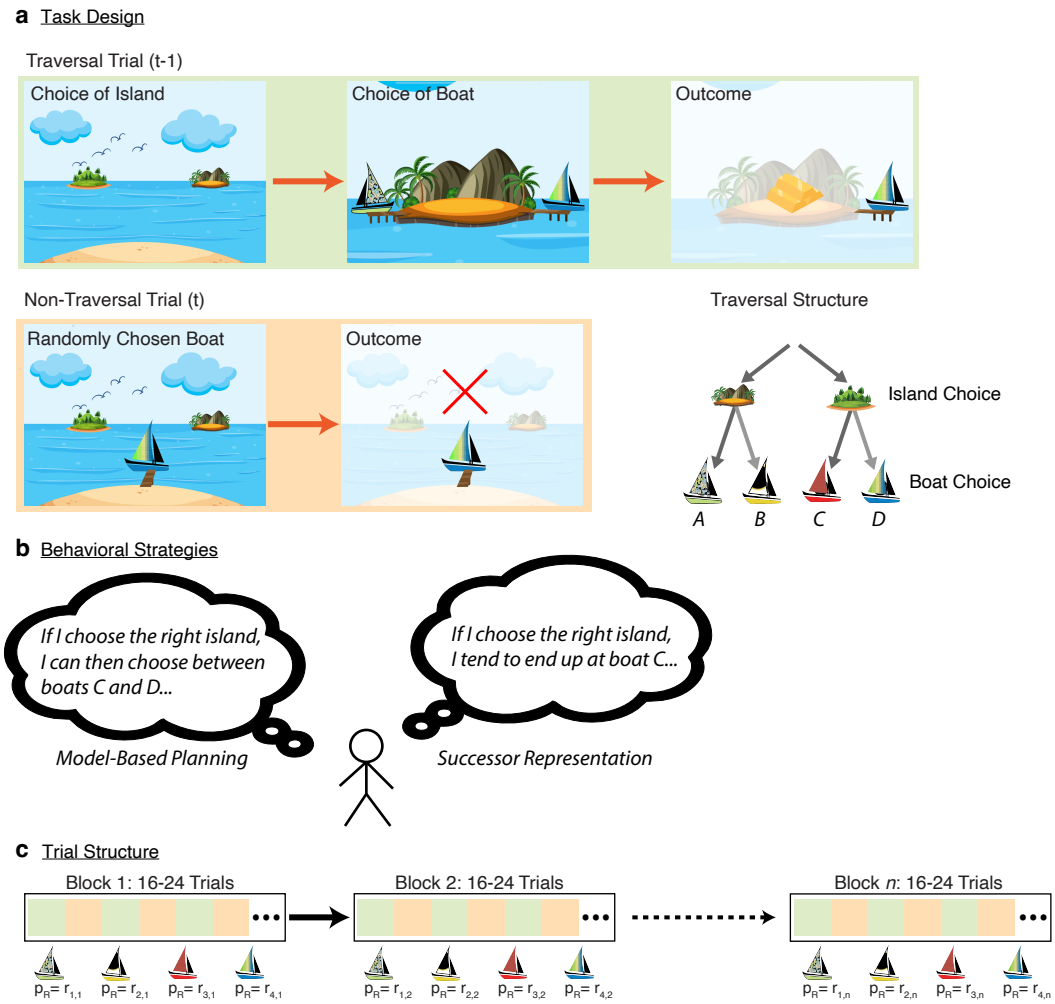

Figure 1: **Planning Model and Task Structure**. (**a**) Trials alternate between "traversal" trials (top) where subjects choose an island and then a boat, and "non-traversal" trials (bottom) in which a boat is selected at random and its reward delivered, without island choice or presentation. On each traversal trial, the participant selects either the left or right island. Upon selecting an island, the two boats available on that island appear, and the participant selects either the left or right boat. On each non-traversal trial, the participant is not given the option to select an island or boat, and instead, one of the four boats "visits them" at the starting location, with identical payout probabilities. The locations of all boats are fixed for the duration of the task. Bottom Right: Full schematic of trial structure. (**b**) A MB agent (left) evaluates choices via recursive maximization, while an SR agent (right) evaluates choices via on-policy averaging. Both SR and MB agents are capable of integrating nonlocal information through the use of an appropriate world model. (**c**) The experiment consisted of 22 reward blocks of between 16-24 trials, alternating between traversal and non-traversal trials. Reward probabilities were consistent within a given block, and systematically altered between blocks.

(and, in turn, islands) most likely to provide treasure. The task was framed as a treasure-hunting game, consisting of 200 pairs of alternating 'traversal' and 'non-traversal' trials. Traversal trials

allowed subjects to navigate to one of two islands, and then one of two boats (Fig. 1a, top), whereas non-traversal trials involved no decision making, with one of the boats visited at random (Fig. 1a, bottom left). Thus, non-traversal trials provided information about the value of boats, but without explicitly associating that information with the path to arrive at that boat. Importantly (compared, for instance, to other two-step Markov decision tasks), this ensured that any effect of this information on the next island choice was mediated via some type of a world model (i.e., it ruled out MF learning about the island choice on these trials), enabling a direct comparison of MB vs. SR approaches to this type of model-based credit assignment (Fig. 1b).

Participants were instructed that reward probabilities were properties of the boats alone, independent from one another and from the particular action sequences the subjects might take to achieve them (including being the same on traversal vs. non-traversal trials), but might change over the course of the game. Nonetheless, rewards did follow certain patterns, unknown to the subjects: reward probabilities were fixed for blocks of 16-24 trials (or 8-12 pairs of traversal and non-traversal trials), after which they would switch to a new set of reward probabilities for the subsequent block (Fig. 1c). Of note, the probabilities were structured so that at all times both islands had an equal probability of reward, if boats were chosen at random, so that non-traversal trials were equally likely to result in reward attributable to either island.

Participants Use World Models to Incorporate Reward Information

The task was designed to distinguish between MB and SR strategies on a per-trial basis, because these predict different characteristic patterns by which reward information from non-traversal trials would be incorporated into subsequent decisions. We can understand and provide an initial test of the models' predictions by focusing specifically on the effect of each non-traversal trial's reward on the island choice made on the subsequent traversal trial. These types of simplified, one-trial-back analyses help complement and provide intuition about the factors driving the relative fit of more elaborate learning models, which we consider later. First, to verify that participants were indeed incorporating reward information from these trials, we examined whether receiving reward (vs. not) for some boat on a non-traversal trial affected participants' chance of choosing the island associated with the sampled boat on the immediately following traversal trial. Both approaches to world modeling (MB and SR) predict a win-stay-lose-shift pattern, on average, at the island level: that is, to the extent they follow either strategy, participants should be more likely to choose the associated island if the boat was rewarded rather than unrewarded on the immediately succeeding trial.

Indeed, participants were more likely to choose the associated island when the previous non-traversal trial was rewarded than when it was unrewarded (effect of reward: $\beta = 0.5, z = 9.66$, two-sided $p < 1e-21$; mixed effects logistic regression), indicating a baseline sensitivity to reward (Fig. 2a). (Note that this effect cannot be ascribed to subjects being anyway more likely to prefer islands associated with rewarded boats, since the probability the sampled boat is rewarded is at all times matched across islands.) Importantly, since the choice of an island was not experienced on a non-traversal trial, this effect implies the use of some type of internal model to infer the implication of a boat's reward for the choice of islands. This allowed us to further decompose this reward effect, to investigate to what extent it reflected the contributions of either SR or MB planning mechanisms.

People Exhibit Signatures of Both SR and MB Evaluation

Next, we sought to distinguish whether sensitivity to reward (here, shown as the difference between the rewarded and unrewarded conditions, Fig. 2a, left) could be ascribed to MB or SR planning. Both of these mechanisms use (updated) per-boat reward estimates (denoted r_1, r_2 etc.) to construct

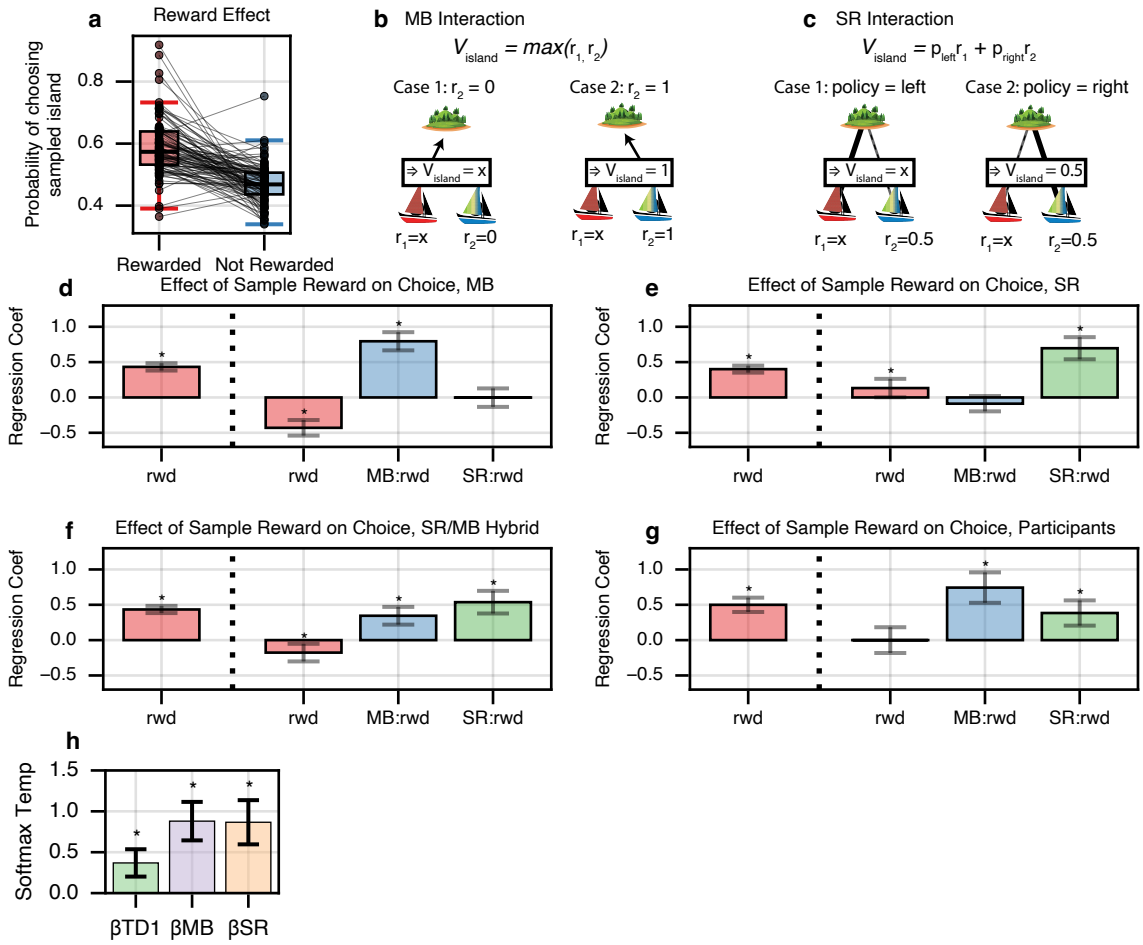

Figure 2: **Distinct Signatures of MB and SR**. (**a**) Reward Sensitivity: Participants are significantly more likely to choose the island associated with a sampled boat when the boat was rewarded, behavior which cannot be explained by MF planning mechanisms. (**b**) MB signature: A MB agent maximizes over boat values at an island, so it is only sensitive to a boat's reward (or nonreward), here indicated as 'x', if the neighboring boat has low value (recently unrewarded). (**c**) SR signature: An SR agent predicts value from expected boat occupancy, which depends on previous choices at the island and is thus only sensitive to a boat's reward if that boat was chosen previously. (**d-g**) Effect of non-traversal reward on subsequent traversal island choice. Left-most bar: Reward-only model. All four models show a significant effect of modulation due to reward on the previous trial. Right three bars: Model including MB and SR interaction terms. Effect of reward almost entirely disappears when interactions with neighboring boat and prior policy are included. MB learners show a strong interaction with the neighboring boat, whereas SR learners show a strong interaction with prior choice at the sampled island. Hybrid agents, as well as humans, exhibit both interaction effects. (**h**) Estimated group-level softmax temperatures for TD1, MB and SR strategies. Error bars are 95% CI.

(updated) values for each island, but they do so via different processes that, in effect, embody different assumptions about which boat will be chosen at an island. These, in turn, mean that reward feedback about a boat will have more or less impact on its island's value in different circumstances,

which we can detect by testing whether the reward effect described above is further mediated by different interacting variables.

Full MB planning iteratively recomputes the value of the maximizing choice at each step (i.e. evaluates the island under the assumption that the currently best boat is chosen there). This means in our task that sensitivity to feedback about a boat's reward should be effectively gated by the value of the island's alternative boat: in particular, any feedback about *Boat* 1's reward should have less effect on the MB-estimated value $V_{island} = max(r_1, r_2)$ if the neighboring *Boat* 2 is already estimated to be near the ceiling value of one (Fig. 2b). Conversely, the SR simplifies this iterative, nonlinear computation to a single linear one, by averaging each state's reward according to its predicted occupancy, which in turn the SR learns from previous state encounters rather than infers by maximizing over the (updated) state values. In the current task, such previous encounters with boats are governed by the subject's own pre-existing choice policy at the island: thus, $Boat\ 1's$ updated contribution to the SR-computed value $V_island = p_{left}r_1 + p_{right}r_2$ is weighted by p_{left}, the chance of choosing the boat prior to receiving the new information (Fig. 2c).

Accordingly, the two models predict that the impact of an update to *Boat* 1's value is modulated by different factors, motivating a test for either signature by interacting the reward with different variables. We can test these qualitative signatures in a model-agnostic logistic regression by testing whether the next-trial reward effect (Fig. 2a) interacts with the gating variables (r_2 and p_1, respectively). Since these are themselves covert variables within the models, for a model-agnostic analysis, we define simple, objective proxies for them in terms of individual recent events (these also represent simple limiting cases of the full computational learning model for high learning rates). In particular, we approximated the neighbor value r_2 by the reward (0 or 1) received the last time that boat was encountered; and we similarly approximate the default choice policy p_1 by a binary indicator (0 or 1) for whether *Boat* 1 was chosen the last time its island was encountered.

Fig. 2d-f demonstrates these effects on simulated choice data for 200 agents simulated with either MB planning (Fig. 2d), SR planning (Fig. 2e) or an intermediate mixture of both MB and SR (Fig. 2f). To verify the robustness of the one-trial-back analysis when learning is actually incremental over time, these agents used learning rates similar to those estimated from fits to our human data, and the regressions also contain additional nuisance variables to control for potential bias due to these longer-term effects (see Methods). Using a hierarchical regression model, we first verified that all agents showed a main effect of a non-traversal-trial reward on the next traversal-trial island choice (left-most bar in all plots). Next, we extended this model to include interaction terms with (our proxies for) the neighboring boat value (MB) and the prior policy (SR) (along with nuisance variables such as lagged effects of these interactions on earlier trials, see Methods). These analyses verified that these interactions were each selectively significant for the simulations of the corresponding model regime, while both signatures coexisted for the intermediate case. Of note, these interacting effects largely explain away the main effect of reward seen in the model when the interactions are not considered: the beta coefficient for reward is drastically diminished or even negative in all three cases.

We next fit the same pair of models to human participants (Fig. 2g), and find that similar to a mixture of MB and SR planning, participants exhibited a sensitivity to reward from non-traversal trials ($\beta = 0.5$, 95% CI = +/- 0.101, $p < 4.61e-22$) which was entirely mediated by the two interaction terms ($\beta = 0.0002$, 95% CI = +/- 0.182, $p < 0.998$), whereas both interactions with reward show significant positive effects (MB effect: $\beta = 0.743$, 95% CI = +/- 0.215, $p < 1.32e-11$, SR effect: $\beta = 0.384$, 95% CI = +/- 0.179, $p < 2.53e-5$). Notably, this mixture of MB and SR behavior is in-line with previous work such as Momennejad *et al.* [20], which also suggested (using a much different task and blockwise manipulation) that human behavior could be explained with a mixture of the SR and MB planning, but not with either individually. Likewise, we observed in our

task that human behavior, at the group level, displayed a similar mix of behaviors: people rely on both SR and MB evaluations.

Mixture of agents model

While the above model-agnostic analysis helps to provide evidence and intuition for the separability of MB and SR behavioral signatures, actual subject choices depend on expectations learned incrementally over multiple trials. Thus to complement this model-agnostic analysis and further verify that human choice behavior was best-explained by a mixture of MB and SR planning, we fit subject behavior with a fuller learning model, extending the hybrid "mixture of agents" model often used previously in sequential decision tasks[29]. Previous models typically contained two agents: MB alongside a model free (temporal difference, TD, learning) agent. Here we added to this mix a third, SR agent. Note that since the MF agent learns experientially, it cannot directly account for reward effects on non-traversal trials; but it might still contribute to choice via preferences learned on traversal trials, whose influence its inclusion accounts for. The three agents' contributions to choice are weighted by separate softmax temperatures (β_{MB}, β_{SR}, β_{TD}), which measure their respective contribution to choice.

The agents shared learning rate parameters. In particular, previous work suggests that people may employ different learning rates for active versus passive learning[33,34], and so we allowed for learning rate to vary between the two trial types: α_A on traversal trials, and α_P on non-traversal trials (active and passively, respectively).

Fits of this model indicated human choice behavior exhibited significant signatures of all three agents, with group-level estimates of the softmax temperatures of all three agents significantly different from zero (Fig. 2h) ($\beta_{\text{TD}} = 0.369$, 95% CI = +/- $0.167, t_{810} = 4.328, p < 1e{-}4$, $\beta_{\text{MB}} = 0.88$, 95% CI = +/- $0.234, t_{810} = 7.33, p < 1e-10$, $\beta_{\text{SR}} = 0.866$, 95% CI = +/- $0.27, t_{810} = 6.282, p < 1e-9$). Consistent with the model-agnostic regressions, this result indicates that human behavior exhibits distinct characteristics of MB and SR planning, even with additional MF contributions taken into account, and that choice behavior is best-fit by a model that allows for a mixture of the three.

Within-Subject Arbitration Between SR and MB strategies

Given the signatures of both SR and MB evaluation, we next wished to investigate how participants balanced these strategies. In particular, extending previous ideas about MB vs. MF arbitration[6–8], we hypothesized that participants should tend to adopt an SR strategy when the simplification underlying it was more often valid, and fall back on more expensive MB computation when the SR's assumptions were more often violated. In particular, the SR simplifies iterative prediction by assuming that expectations about state encounters (here, boats given islands) are consistent over time: in our task, this requires that participants' own per-island boat preferences are stable. To test this, our task manipulated the validity of this strategy through systematic blockwise switches in reward probability which induced more or less policy instability. The rewards associated with each of the four boats underwent frequent switches (Fig. 3a).

In particular, at the beginning of the experiment, reward probabilities were initialized such that each island had a more and a less rewarding boat, and the better boat in one island was in turn better than the other island's (better $p(reward) = 0.85$ vs. 0.675). Every 16-24 trials, reward probabilities would reverse in one of two ways, either favoring or disfavoring the SR. First, the rewards might undergo a *congruent change*, (15/21 block changes), in which probabilities were swapped between the two higher boats and between the two lower boats, such that optimal choice

of island changed, but optimal choice of boat conditioned on island remained the same (Fig. 3a, left). Second, rewards might undergo an *incongruent change, 6/21 block changes*, which changed the optimal boat choice at each island (Fig. 3a, right). (Half of these also swapped the optimal island choice; half did not.) The key observation is that the incongruent block changes violate the assumptions of the SR: island choices based on the SR will mispredict the island's optimal value (recommending suboptimal island choices), and to the extent participants relearn new boat choices, the SR's predictions about boat encounters will initially be violated, and the SR will require updating. We hypothesized that the brain could detect these effects (e.g., via prediction errors on boat encounters) and adaptively reduce SR usage following incongruent block shifts. (We included fewer incongruent than congruent reversals based on the expectation, confirmed in piloting, that selective adaptation would be most likely elicited by occasional rather than frequent challenges.)

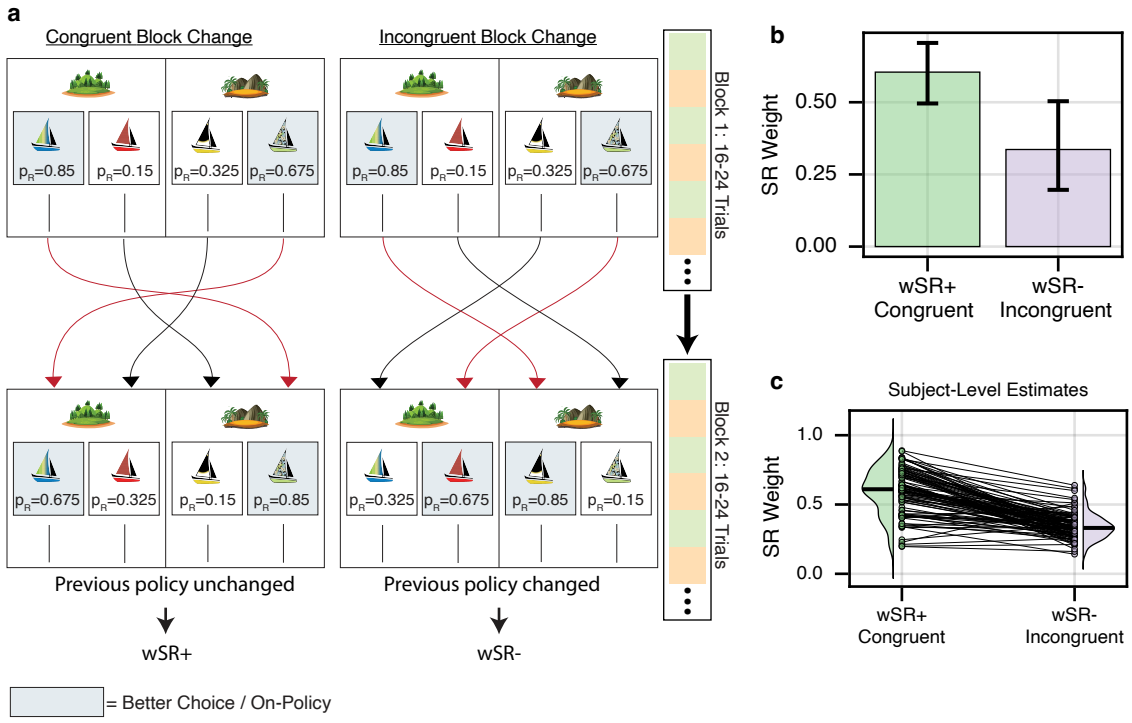

Figure 3: **MB/SR balance systematically varies across block reward structure**. (**a**) Left: Congruent block change. After 16-24 trials, reward probabilities swapped such that the optimal choice of island flipped, but the optimal boat choice conditioned on island choice remained unchanged. Crucially, evaluating the new rewards via the policy induced by previous rewards still provided the optimal choice of island. Right: Incongruent block change. Again, the optimal island changed, but now, the policy induced by previous rewards valued the sub-optimal island more highly. (**b**) Parameter fitting results. A hierarchical choice model was fit across subjects which included a weighting term w_{SR} parameters controlling the balance between MB and SR behavior. *wsrp* was used on blocks following congruent reward changes, whereas *wsrm* was used on blocks following incongruent reward changes. *wsrp* was significantly greater than *wsrm* at the group level (two-sided $t(1044) = 2.797, p < 0.006$), suggesting that subjects increasingly relied on the SR when future state occupancy was stable, and MB planning otherwise. Error bars show +/- 1.96 SE. (**c**) Individual subject fits. Each line is one subject.

To estimate whether subjects' balance of MB and SR planning systematically varied across blocks, we again fit the mixture of agents model to choice behavior, but now estimating separate MB, SR, and TD1 softmax temperatures for each block type, determined by which type of reversal began that block. We predicted that when subjects encountered new reward probabilities under which optimal choice of island was consistent with prior policy, participants would rely more heavily on their prior policy, and their behavior would be best estimated with a larger relative reliance on SR over MB, reflected in higher SR and/or lower MB softmax temperatures. If, instead, new reward values led to incorrect choice of islands under the prior policy, we expected that participants would spend relatively more effort on explicit MB planning, which would be reflected in increased MB and/or decreased SR softmax temperatures. To formally test for blockwise changes in relative SR vs. MB reliance, we computed their fractional contribution $w_{\mathrm{SR}} = \beta_{\mathrm{SR}}/(\beta_{\mathrm{SR}} + \beta_{\mathrm{MB}})$, such that a weight of 0 indicated complete reliance on MB and a weight of 1 complete reliance on SR[29]. (This was accomplished by estimating the original model using a change of variables; see Methods.)

Estimating the model hierarchically across subjects, we found that the group-level mean $w_{\mathrm{SR}}+$ (the reliance on SR relative to MB planning following a congruent reward change) was, as hypothesized, significantly larger than $w_{\mathrm{SR}}-$ (the value following an incongruent reward change). Following a congruent reward change, $w_{\mathrm{SR}}+$ was estimated at 0.604 (95% CI: 0.496 to 0.705) indicating an SR-dominant strategy. Following an incongruent reward change, $w_{\mathrm{SR}}-$ was estimated at 0.336 (95% CI: 0.197 to 0.504), indicating a shift to an MB-dominant strategy. This change was itself statistically significant (t(1044)=2.797, p < 0.006). In contrast, there was no significant blockwise difference in the MF contribution, ($\beta_{\mathrm{TD}}+ = 0.236$, 95% CI $= [0.066, 0.406]$, $\beta_{\mathrm{TD}}- = 0.156$, 95% CI $= [-0.018, 0.330]$, estimated difference $= 0.079$, 95% CI $= [-0.136, 0.294], t(1044) = 0.720, p <$ 0.47), nor in the overall MB/SR softmax temperature β_{MBSR} that, on this change of variables, measures the contribution of a w_{SR}-weighted average of MB and SR values ($\beta_{\mathrm{MBSR}}+ = 2.029$, 95% CI $= [1.669, 2.388]$, $\beta_{\mathrm{MBSR}}- = 1.637$, 95% CI $= [1.274, 2.001]$, estimated difference $= 0.392$, 95% CI $= [-0.075, 0.859], t(1044) = 1.644, p < 0.11$).

These results indicate that on blocks following congruent reward changes, subjects exhibited increased reliance on temporally abstract planning, whereas following incongruent reward changes, subjects turned more to MB evaluation. This difference was also relatively consistent in the hierarchical model's subject-level estimates: 93/100 participants were individually estimated with $w_{\mathrm{SR}}+ > w_{\mathrm{SR}}-$. Thus we observe that people are capable of reacting to the temporal predictability of their current environment, and dynamically adjusting their decision-making strategy to balance cost-of-planning against expected outcome.

We further confirmed that this difference in estimated parameters was not induced by the blockwise structure of the experiment itself: when simulating an agent insensitive to the blockwise structure and recovering softmax temperatures for each block, we found that there was not a significant estimated shift between block types (see Supplement).

Linear RL

Although we have so far estimated SR and MB contributions via separate modules in a standard mixture-of-agents model, recent theoretical work suggests an alternative formulation in which both algorithms arise as special cases of a single world model computation with tuneable reliance on temporal abstraction [30, 35]. Though largely behaviorally equivalent, this suggests a different view on the neural implementation of world models and the cost of MB planning.

In particular, a linear RL agent computes values under a default assumed dynamics (equivalent to SR's M), but can partly correct for this and approximate the iterative maximization underlying MB choice using a softmax-like nonlinearity [30, 35]. The degree of this correction is controlled

by a single scale factor, λ, which, like w_{SR}, parameterizes a spectrum of behavior from SR to MB (Fig. 4a). Furthermore, this parameterized scaling implies a concrete computational cost in terms of bits of precision (e.g. more spikes), as better approximations to MB choice require computing larger, more precise values: (Fig. 4b). As expected (since the model is largely behaviorally equivalent on the current task), this alternative specification, when estimated on our data, recapitulated our core results. In particular λ was overall estimated as intermediate (between extreme values characteristic of pure SR or MB) (Fig. 4c), and declined (nearer to MB) on incongruent blocks when compared to congruent blocks (Fig. 4d,e).

Discussion

When do we spend time and mental resources iteratively simulating the future, and when do we rely on simpler procedures? Simplified evaluation strategies inherently trade off planning complexity against accuracy, and the usefulness of any given approach for planning depends on the environment an agent is operating within[6]. This has often been discussed in terms of a tradeoff between model-based and model-free methods. MF learning is advantageous when the long-run values of candidate actions change slowly enough for a cached estimate to accurately reflect distant reward outcomes[32]. It has been hypothesized that the brain trades off such strategies judiciously according to circumstance[36], and (in the case of MB vs. MF arbitration) that this process explains phenomena of both healthy automaticity and pathological compulsion[6,12,37,38].

Here we extended this program to the question of how the world model itself works. Simplified world models, such as the SR, have seen recent attention in decision neuroscience. These simplify planning not by foregoing model-based simulation altogether, but instead by collapsing the iterative simulation of dynamics by using temporal abstraction over future state encounters: in effect, adopting a simplified, multi-step world model[18,19,30]. Furthermore, one candidate implementation of full MB planning in the brain involves an inherent, tuneably cost-saving tradeoff with SR-like abstraction, suggesting that judicious MB-SR tradeoffs may be as, or even more, fundamental than the classic MB-MF picture[30]. The SR and related methods provide a potential solution to how the brain can plan efficiently over temporally extended timescales, so long as the associated predictions are stable, but also predict subtler biases. We introduced a new task and analyses that allowed us to track SR vs. MB usage dynamically in terms of trial-by-trial choice adjustments rather than, as in previous work, via one-shot probes[20]. Using these methods, we confirmed previous suggestions that people use such temporally abstract models (the SR) alongside iterative MB planning mechanisms. This allowed us further to investigate dynamic strategy usage within-participant, providing evidence that people dynamically adjust the balance between these strategies in response to their accuracy.

Dynamic Adaptation

Previous research on arbitration between MF and MB planning systems suggests that people dynamically adjust their behavioral strategies, optimizing for a cost-benefit trade-off[7,8] between fast but noisy estimates and slow deliberation. For example, when faced with frequently shifting long-run outcomes, people should increasing rely on MB strategies, despite the additional cognitive cost[32]. That is, the brain is capable of recognizing when cached value estimates are likely to be unreliable, and spends additional time and effort to compensate. Here, we showed that the same is empirically true of SR and MB planning: when cached policy estimates were unreliable vis-a-vis optimal rewards, people shifted towards MB strategies.

What causes a shift in reliance on the SR versus MB planning? Prior work on MF behavior suggests that people may rely either on prediction accuracy[6], or on estimates of sub-optimality of

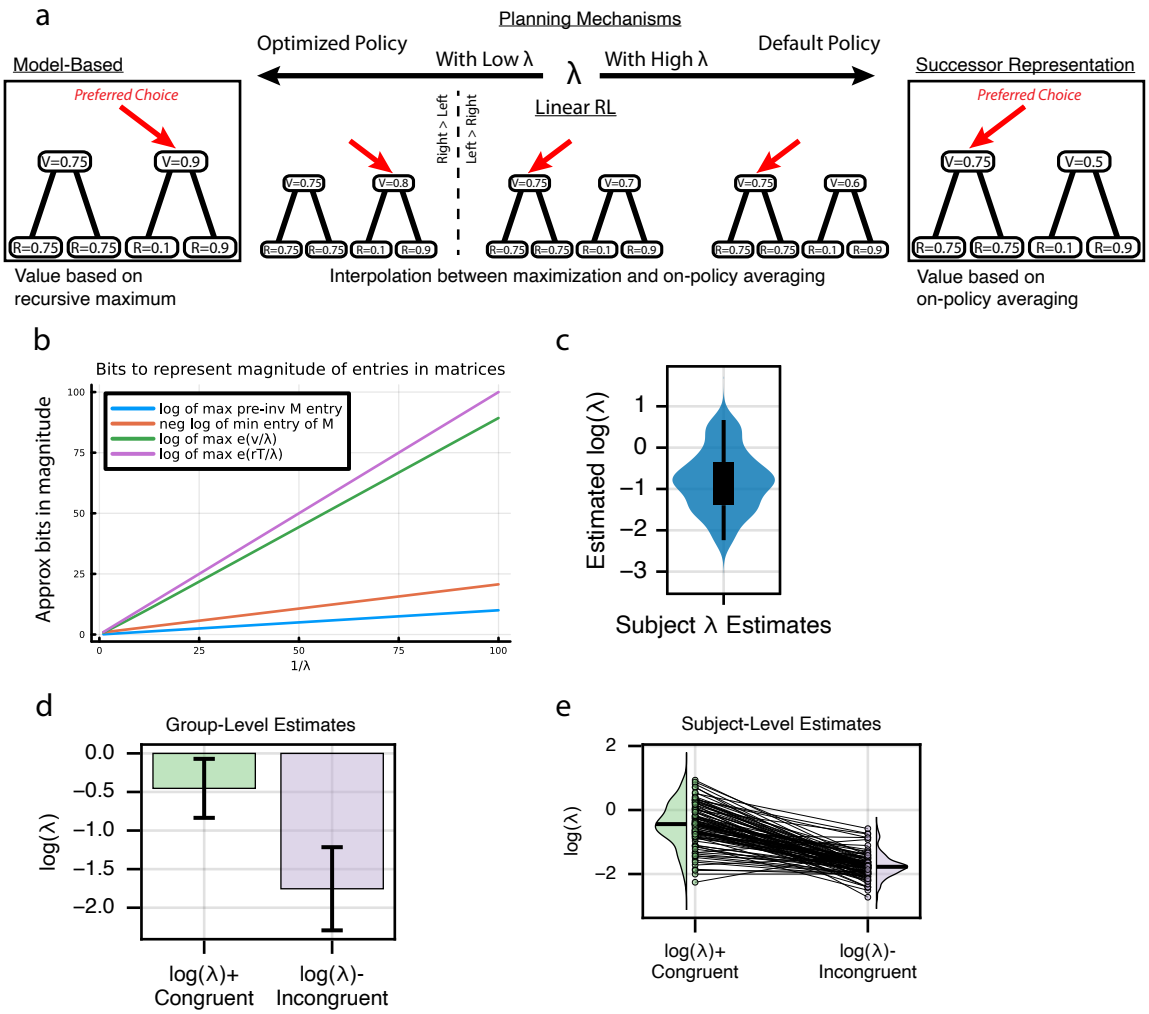

Figure 4: **Linear RL λ systematically varies across block reward structure**. (**a**) Linear RL encapsulates both MB and SR behavior. As a single parameter λ varies between extremes, linear RL approaches either MB (low λ) or SR (high λ) evaluation. (**b**) Tuning for MB behavior exhibits an inherent cost. As λ approaches zero, the necessary computational precision dramatically increases, suggesting a fundamental neural cost to increasingly accurate read-outs. Here, the relationship between the magnitude of the largest value and the inverse of λ is shown for the current task. (**c**) Parameter fitting results. A hierarchical choice model was fit across subjects which included a single λ parameter controlling the balance between MB and SR behavior. $log(\lambda)$ was estimated at a value of -0.803, 95% CI $= +/-0.323$, suggesting that subjects rely on a mixture of MB and SR-like behavior (see Supplement). (**d**) Blockwise fits. A hierarchical choice model was fit across subjects which included two λ parameters controlling the balance between MB and SR behavior. $\lambda+$ was used on blocks following congruent reward changes, whereas $\lambda-$ was used on blocks following incongruent reward changes. $log(\lambda+)$ was significantly greater than $log(\lambda-)$ at the group level (two-sided $t_{728} = 3.629, p < 0.0004$), suggesting that subjects increasingly relied on the SR when future state occupancy was stable, and MB planning when unstable. Error bars show +/- 1.96 SE. (**e**) Individual subject fits. Each line is one subject.

reward obtained under the current policy[8]. Here, our experimental design exhibited both of these characteristics at the block level: incongruent blocks led to both more frequent state prediction errors, as well as increased losses if islands were chosen via the previous policy. Yet another possible signal is whether outcomes are generally consistent with the future occupancy expected under the SR — if rewards tend to occur in states on which we put little future expected weight, or vice-versa, we may infer that the current policy is invalid. While a long-term solution is to relearn updated state occupancy estimates, a short-term reaction may be to temporarily reduce reliance on the SR, and instead perform more costly MB searches. The reliance on a blockwise measure of arbitration in the current study allows us to capture coarse behavioral shifts, but future work will be required to test which precise characteristics drive these shifts, including the use of finer-grained models which these these hypothesized mechanisms of trial-by-trial adaptation against the current behavioral paradigm.

Trial-by-Trial Dynamics of Temporal Abstraction

Our work builds on the logic of previous approaches to experimentally distinguish SR-based from step-by-step MB planning. One key signature of the SR is a limited ability to successfully replan when distant rewards change: replanning is hampered (and step-by-step MB re-evaluation required) when reward changes also affect the preferred choice policy at subsequent states[19]. To test for this limitation, previous approaches have relied on one-shot train-retrain-probe designs ("policy revaluation" retraining), contrasting one-shot change scenarios in which an SR planner should vs. should not be able to successfully replan. Because of their binary nature (and susceptibility to practice effects if repeated), these provide limited information about individual differences, instead only providing a generalized summary of human behavior on average over subjects[20]. Relatedly, this class of designs is also not ideal for extended training in animals, nor for dynamic neuroscientific measurements and manipulations, like neuroimaging, electrophysiology, or optogenetics.

Using a novel choice task, we extend the logic of policy revaluation to the per-trial effects of individual rewards on trial-by-trial choice adjustment. Specifically, the reliance of MB planning on successive maximization and of SR on the previously experienced policy both lead to distinct effects of updated reward information, when that reward information is interspersed between trials probing the subject's action preferences. In effect, in this task each "non-traversal" trial is a miniature version of a (policy and/or reward) revaluation training phase. By measuring these effects we can measure the influence of these interactions over the course of the experiment. Extending previous approaches that used similar logic and designs to distinguish MB from MF updating[29], we show that trial-by-trial learning dynamics provide a rich source of information about the details of internal planning computations: in effect, how participants model the world's dynamics for the purpose of evaluating choices.

Indeed, similar to the large body of previous work using the two-step Markov decision task to distinguish MB from MF, we expect that the ability to measure SR usage dynamically will analogously enable informative future studies. These might include the ability to measure and manipulate neural correlates of these planning processes dynamically during learning, to further investigate arbitration driven by cost-benefit trade-offs between multiple planning systems[7,8], and, via better characterization of individual differences in these behaviors, to investigate whether usage of SR systematically varies in psychiatric conditions (much as has been argued for MF)[12], as well as over the course of normal human development[15].

Trade-offs in Linear RL

Traditionally, phenomena of automaticity vs. deliberation have been viewed in terms of MF vs. MB competition: retrieving cached long-run values vs. updating them via model search[6]. The SR, in turn, was first conceived as a third, intermediate strategy, which flattens and thus more cheaply approximates model search by caching long-run trajectories[18,19]. This is indeed the structure of the descriptive, mixture-of-agents model by which we primarily analyze our data.

However, recent theoretical work on linear RL[30] suggests an alternative implementation of MB computation in the brain, in which a tradeoff against SR-like simplification occurs intrinsically within a single computation. In linear RL, values are predicted under an assumed multistep dynamics, similar to the SR, but with an additional softmax-like nonlinearity that can, in the limit, mimic full MB computation. This tradeoff, in turn, has a concrete cost (e.g. in terms of spikes or time): more MB-like values require larger, more precise decision variables. On this view, then, tuneable MB vs. SR tradeoffs are inherent to word modeling, and potentially as fundamental as MB vs. MF. This is thus a natural framework for considering rational meta-control of the sort we find evidence for here.

That said, although this is an appealing perspective and motivates the current study, our experiment was not designed to distinguish linear RL from a mixture of separate SR and MB agents – both of which suffice, at a descriptive level to capture our key findings. Thus, our results do not yet speak directly to whether the underlying operations, and their evidently rational balancing, are implemented in a single or separate neural "modules."

Acknowledgments

This work was supported by grant 61454 from the John Templeton Foundation and grant W911NF-16-1-0474 from the U.S. Army Research Office. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

Author Contributions

A.E.K. and N.D.D. conceived the project and planned the experiments and analyses. A.E.K. performed the experiments and analyses. A.E.K. and N.D.D. wrote the manuscript.

Declaration of Interests

The authors declare no competing interests.

Methods

Participants. A total of 104 participants were recruited through Prolific to participate in an online behavioural experiment in September 2023. This study was approved by the Institutional Review Board of Princeton University, and all participants provided informed consent. The total study duration was approximately forty minutes per participant. The participants received monetary compensation for their time (rate US$12 per hour), plus a task performance bonus up to $3.00 based on treasure collected in-game.

Participants were eligible if they were fluent in English, resided in Australia, Canada, New Zealand, or the United States, and had not participated in any prior versions or pilots of the study. Subject exclusion critera were implemented through Prolific.

Code. The task was implemented using NivTurk[39] and jsPsych[40] and distributed using custom web-application software. All experiment code is publicly available. A playable demo of the task is available at `https://ariekahn.github.io/policy-sailing-task`.

Experimental Task. Participants performed a task where they repeatedly chose among four merchant boats, each with a hidden probability of reward, in order to maximize their monetary payout (a bonus of up to $3.00) by selecting the boats most likely to provide reward. The task was framed as a treasure-hunting game, consisting of 200 sets of alternating 'traversal' and 'non-traversal' trials. Participants were instructed that each merchant had a probability of reward independent of the other merchants, that these probabilities might change over the course of the game, that visiting a merchant did not increase or decrease future probability of reward, and that navigating to a merchant vs. having the same merchant visit the participant had the same probability of reward. The visual appearance of boats and islands was shuffled across subjects, as was the pattern of reward probabilities (see **Reward Block Structure**).

Trial Structure. All trials started from the same 'home' island. On each 'traversal' trial, the game presented two candidate islands, and the participant chose to sail from their home island to either the left or right island by pressing the corresponding arrow on the keyboard. The game would then zoom in on the selected island, and two merchant boats would appear on the left or right side of the island, which incorporated a 1s delay, preventing the participant from entering their second-stage choice until this transition completed. The participant was now provided with a second choice between the two boats, and upon selecting a boat with the left or right arrow, it was revealed whether the chosen merchant would provide treasure - either a stack of gold bars (+1) or a red 'X' (+0) was shown on the screen. Notably, the position of the islands as well as the boats was constant throughout the experiment, though boats were not visible on the screen until the corresponding island was first chosen. On a 'non-traversal' trial, one of the four boats, at random, was shown to the participant at their home island. The participant pressed 'up' to select the boat and reveal whether the merchant would provide treasure. Each boat appeared 25 times, with a randomized order across the 200 'non-traversal' trials. Probability of reward was identical to the probability of reward for that boat on the preceding 'traversal' trial. In effect, these trials served to provide information about how likely one boat was to have reward, independent of the participant's own policy.

Reward Block Structure. The experiment consisted of 22 reward blocks of between 8 to 12 trial pairs. Reward probabilities were consistent within a given block, and systematically altered between blocks. At the start of the first block, one island consisted of merchants with probabilities of reward of 0.15 and 0.85, and the other island of merchants with probabilities of reward of 0.325 and 0.675. Assignment to each boat was randomized for each participant.

On a policy-consistent ('congruent') swap, the probabilities of the optimal boats on each island were swapped with one another, as were the probabilities of the sub-optimal boats. On a policy-inconsistent ('incongruent') swap, the probability of the optimal boat on one island was swapped with the probability of the sub-optimal boat on the other island. On a within-island swap (also 'incongruent'), the reward probability of each boat was swapped with that of the other boat on the same island. Suppose the boats are given by (A1, A2) on one island, and (B1, B2) on the other, with initial reward probabilities of (A1, A2, B1, B2) = [0.15, 0.85, 0.325, 0.675]. Then a policy-consistent swap would result in [0.325, 0.675, 0.15, 0.85], a policy-inconsistent swap would result in [0.625, 0.325, 0.85, 0.15], and a within-island swap would result in [0.85, 0.15, 0.675, 0.325]. These particular re-assignments were chosen such that in a congruent swap, a policy optimized for the previous reward structure will still correctly identify which island is better under the new rewards, but for an incongruent swap, a policy optimized for the previous reward structure will *incorrectly* identify which island is better under the new reward structure.

Data Exclusion. Participants whose mean reaction time was more than three standard deviations from the mean on either boat or island trials were excluded (1 outlier on island RT, 3 outliers on boat RT) leaving 100 subjects.

Regression Model

To test for a general effect of reward, behavior was fit to a hierarchical generalized linear mixed model using *MixedModels* v4.22.1[41] in *Julia* 1.9.3[42] with a Logit link function and the following formula:

$$actionTowardsSample \sim 1 + rwd \tag{1}$$

where *actionTowardsSample* is whether the participant chooses the island associated with the sampled boat, and *rwd* is whether the sampled boat was rewarded. Per-subject random effects were included for both slope and intercept.

To test for interactions representative of either policy-based or model-based credit assignment, behavior was fit to a hierarchical generalized linear mixed with a Logit link function and the following formula:

$$\begin{aligned} actionTowardsSample \sim\ & 1 + rwd + sibNoRwd + onPolicy \\ & + rwd\ \&\ sibNoRwd + rwd\ \&\ onPolicy \end{aligned} \tag{2}$$

where *actionTowardsSample* is whether the participant chooses the island associated with the sampled boat, *rwd* is whether the sampled boat was rewarded, *sibNoRwd* is whether the most recently observed outcome from the second boat on the same island was unrewarded, and *onPolicy* is whether the previous action from the sampled boat's island was to the sampled boat. In addition, to control for non-independence between trials due to effects of events on previous trials, we added a number of lagged regressors: specifically, we included an additional four outcomes on the neighboring boat, four prior choices at the same island representing policy, the interactions of those with the reward of the previous trial, the previous five outcomes and interactions with reward of outcomes at the sampled boat, five lags of the average of the value of the two boats on the opposite island, and five lags of an island choice kernel for whether the participant had chosen the associated island on the last five traversals. All regressors are relative to the boat that was just sampled and the island on which that boat is found, past of the policy, current boat, and neighbor boat were *sibNoRwd* and *onPolicy* coded 0/1, and all other regressors were coded -0.5/0.5.

Behavioral Choice Model

To estimate the expected value of choosing each of the four boats, we used a simple value-learning algorithm, where the value of boat i was updated after an observed reward R_t:

$$V(boat_i) \leftarrow V(boat_i) + \alpha R_t \tag{3}$$

The learning rate differed for traversal and non-traversal trials, to allow for separate active and passive learning rates: $\alpha \leftarrow \alpha_A$ on traversal trials and $\alpha \leftarrow \alpha_P$ on non-traversal trials, with both estimated independently of one another.

Choice between boats was modeled as a probabilistic decision between the two boats present on the chosen island on each trial using a softmax distribution:

$$P(b_t = b \in L, R) \propto \exp(\beta_{boat} V(b) + \beta_{persistenceB} LastChosen(b)) \tag{4}$$

In human parameter estimation (but not simulation) $\beta_{persistenceB}$ was added to control for persistent choice behavior, where $LastChosen(c)$ is 1 if b was the most recently chosen boat at the current island, and 0 otherwise.

To estimate the expected value of choosing the left versus right island using a mixture of agents, values were estimated independently for a MB, SR, and TD(1) agent.

For the MB agent, the value of each island was simply taken as the maximum of the two estimated boat values:

$$V_{\text{MB}}(i) = \max(V(boat_1), V(boat_2)) \tag{5}$$

For the SR agent, the value of each island was calculated via a matrix of future state occupancy, M, where M was itself learned via a hebbian learning rule, and α_M was a learning rate for the M matrix which was fit as a free parameter:

$$M[island, boat] = \alpha_M 1_{boat=c} + (1 - \alpha_M) M[island, boat] \tag{6}$$

Island values were then calculated as $V_{\text{SR}} = MR$, where R is a vector of state rewards.

For the TD agent, island values were updated at the end of each traversal trial as follows:

$$V_{\text{TD}}(island) = (1 - \alpha_A) V_{\text{TD}}(island) + \alpha_A V_{\text{TD}}(boat) \tag{7}$$

Choice between islands was modeled as a probabilistic decision between the two state values, again using a softmax distribution but with a separate inverse temperature for each agent (β_{MB}, β_{SR}, and β_{TD}), as well as persistence estimate $\beta_{persistenceI}$ in human model fitting:

$$P(i_t = i \in L, R) \propto \exp(\beta_{\text{MB}} V_{\text{MB}}(i) + \beta_{\text{SR}} V_{\text{SR}}(i) + \beta_{\text{TD}} V_{\text{TD}}(i) + \beta_{persistenceI} LastChosen(i)) \tag{8}$$

Linear RL Choice Model

To estimate the expected value of choosing the left versus right island using linear RL, values were estimated from a transition matrix $\mathbf{T}$ and rewards $\mathbf{r}$ (both learned over the course of the experiment):

$$\exp(\mathbf{v}_N^*/\lambda) = \mathbf{M}\mathbf{T}_{NT} \exp(\mathbf{r}_T/\lambda) \tag{9}$$

where $\mathbf{v}_N^*$ is a vector of optimal values at non-terminal states, $\mathbf{M}$ is the *default representation* (DM) matrix from all non-terminal to all other non-terminal states, $\mathbf{T}_{NT}$ is the transition matrix from all non-terminal to terminal states, $\mathbf{r}_T$ is the set of rewards (or negative step costs) at all terminal states, and λ arbitrates between exact and policy-guided estimates. Value estimates of the boats were given by taking the logarithm of the left-hand side and multiplying each value by λ.

The DR matrix is in turn given by

$$\mathbf{M} = (\text{diag}(\exp(-\mathbf{r}_N/\lambda)) - \mathbf{T}_{NN})^{-1} \tag{10}$$

where T_{NN} is the transition matrix from all non-terminal to all other non-terminal states and and $\mathbf{r}_N$ is the set of rewards at all non-terminal states.

r was assumed to be 0 at islands, and for boats equal to the estimated V for that boat.

T was updated after each choice using a hebbian learning rule, which was chosen to provide equivalent updates to those of the SR's M matrix:

$$T[island, boat] = \alpha T 1_{boat=c} + (1 - \alpha T) T[island, boat] \tag{11}$$

Choice between islands was modeled as a probabilistic decision between the two state values estimated via linear RL, again using a softmax distribution but with a separate inverse temperature β_{island}, as well as persistence estimate $\beta_{persistenceI}$ in human model fitting:

$$P(i_t = i \in L, R) \propto \exp(\beta_{island} V(i) + \beta_{persistenceI} LastChosen(i)) \tag{12}$$

Parameter Estimation

We optimized the free parameters of the learning algorithms by embedding each of them within a hierarchical model to allow parameters to vary between subjects or simulated sessions. Subject-level parameters were modeled as arising from a population-level Gaussian distribution over subjects. We estimated the model to obtain best fitting subject- and group-level parameters to minimize the negative log likelihood of the data using an expectation-maximization algorithm with a Laplace approximation to the session-level marginal likelihoods in the M-step[43]. For hypothesis testing on population-level parameters, we computed an estimate of the information matrix over the population-level parameters, taking account of the so-called "missing information" due to optimization in the E-step[44], itself approximated using the Hessian of a single Newton-Raphson step.

Blockwise Change Analysis

To estimate whether subjects shifted their strategies after congruent versus incongruent block changes, we fit a two-step choice model parameterized in terms of $\beta_{\mathrm{TD}}+$, $w_{\mathrm{SR}}+$, and $\beta_{\mathrm{MBSR}}+$ (used following a congruent block change), and $\beta_{\mathrm{TD}}-$, $\beta_{\mathrm{SR}}-$, and $\beta_{\mathrm{MBSR}}-$ (used following an incongruent block change), where $\beta_{\mathrm{MB}}+ = (1-w_{\mathrm{SR}}+)\beta_{\mathrm{MBSR}}+$, $\beta_{\mathrm{SR}}+ = (w_{\mathrm{SR}}+)\beta_{\mathrm{MBSR}}+$, $\beta_{\mathrm{MB}}- = (1-w_{\mathrm{SR}}-)\beta_{\mathrm{MBSR}}-$, and $\beta_{\mathrm{SR}}- = (w_{\mathrm{SR}}-)\beta_{\mathrm{MBSR}}-$. This change of variables (analogous to one often used for MB vs. MF tradeoffs; Daw *et al.* [29]) was chosen to allow straightforward comparison and statistical testing between $w_{\mathrm{SR}}+$ and $w_{\mathrm{SR}}-$. Note that again there is no inherent constraint that the congruent block values should be greater or less than the incongruent block values.

For the linear RL model, island values were calculated with linear RL (see above) but with two independent values of λ, termed $\lambda+$ and $\lambda-$, and two softmax temperatures, $\beta_{LRL}+$ and β_{LRL}.. $\lambda+$ and $\beta_{LRL}+$ were used to estimate island values throughout a block following a congruent reward change, whereas $\lambda-$ and $\beta_{LRL}-$ were used to estimate island values throughout a block following an incongruent block change, in addition to $\beta_{\mathrm{TD}}+$ and $\beta_{\mathrm{TD}}-$. Note that again there is inherent constraint that $\lambda+$ should be greater or less than $\lambda-$.

Code availability

The experiment code is available at `https://github.com/ariekahn/policy-sailing-task`. The custom web software for serving online experiments is available at `https://github.com/nivlab/nivturk`. A playable demo of the task is available at `https://ariekahn.github.io/policy-sailing-task`. All analysis code is available at `https://github.com/ariekahn/sailing-paper-code`.

References

1. Uddin, L. Q. Cognitive and Behavioural Flexibility: Neural Mechanisms and Clinical Considerations. *Nature Reviews Neuroscience* **22,** 167–179. ISSN: 1471-0048. doi:10.1038/s41583-021-00428-w (Mar. 2021).
2. Miller, K. J. & Venditto, S. J. C. Multi-Step Planning in the Brain. *Current Opinion in Behavioral Sciences. Computational Cognitive Neuroscience* **38,** 29–39. ISSN: 2352-1546. doi:10.1016/j.cobeha.2020.07.003 (Apr. 2021).
3. Tolman, E. C. Cognitive Maps in Rats and Men. *Psychological Review* **55,** 189–208. ISSN: 1939-1471(Electronic),0033-295X(Print). doi:10.1037/h0061626 (1948).

4. Doll, B. B., Simon, D. A. & Daw, N. D. The Ubiquity of Model-Based Reinforcement Learning. *Current Opinion in Neurobiology. Decision Making* **22,** 1075–1081. ISSN: 0959-4388. doi:10.1016/j.conb.2012.08.003 (Dec. 2012).

5. Sutton, R. S. Learning to Predict by the Methods of Temporal Differences. *Machine Learning* **3,** 9–44. ISSN: 1573-0565. doi:10.1007/BF00115009 (Aug. 1988).

6. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control. *Nature Neuroscience* **8,** 1704–1711. ISSN: 1546-1726. doi:10.1038/nn1560 (Dec. 2005).

7. Keramati, M., Dezfouli, A. & Piray, P. Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. *PLOS Computational Biology* **7.** doi:10.1371/journal.pcbi.1002055 (May 2011).

8. Kool, W., Gershman, S. J. & Cushman, F. A. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science* **28,** 1321–1333. ISSN: 0956-7976. doi:10.1177/0956797617708288 (Sept. 2017).

9. Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M. & Shenhav, A. Expectations of Reward and Efficacy Guide Cognitive Control Allocation. *Nature Communications* **12,** 1030. ISSN: 2041-1723. doi:10.1038/s41467-021-21315-z (Feb. 2021).

10. Brown, V. M., Hallquist, M. N., Frank, M. J. & Dombrovski, A. Y. Humans Adaptively Resolve the Explore-Exploit Dilemma under Cognitive Constraints: Evidence from a Multi-Armed Bandit Task. *Cognition* **229,** 105233. ISSN: 0010-0277. doi:10.1016/j.cognition.2022.105233 (Dec. 2022).

11. Everitt, B. J. & Robbins, T. W. Neural Systems of Reinforcement for Drug Addiction: From Actions to Habits to Compulsion. *Nature Neuroscience* **8,** 1481–1489. ISSN: 1546-1726. doi:10.1038/nn1579 (Nov. 2005).

12. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a Psychiatric Symptom Dimension Related to Deficits in Goal-Directed Control. *eLife* **5** (ed Frank, M. J.) e11305. ISSN: 2050-084X. doi:10.7554/eLife.11305 (Mar. 2016).

13. Brown, V. M. *et al.* Reinforcement Learning Disruptions in Individuals With Depression and Sensitivity to Symptom Change Following Cognitive Behavioral Therapy. *JAMA Psychiatry* **78,** 1113–1122. ISSN: 2168-622X. doi:10.1001/jamapsychiatry.2021.1844 (Oct. 2021).

14. Bolenz, F., Kool, W., Reiter, A. M. & Eppinger, B. Metacontrol of Decision-Making Strategies in Human Aging. *eLife* **8** (eds Büchel, C., Kahnt, T. & Samanez-Larkin, G.) e49154. ISSN: 2050-084X. doi:10.7554/eLife.49154 (Aug. 2019).

15. Nussenbaum, K. & Hartley, C. A. Reinforcement Learning across Development: What Insights Can We Draw from a Decade of Research? *Developmental Cognitive Neuroscience* **40,** 100733. ISSN: 1878-9293. doi:10.1016/j.dcn.2019.100733 (Dec. 2019).

16. Collins, A. G. E. & Cockburn, J. Beyond Dichotomies in Reinforcement Learning. *Nature Reviews Neuroscience* **21,** 576–586. ISSN: 1471-0048. doi:10.1038/s41583-020-0355-6 (Oct. 2020).

17. Botvinick, M. M., Niv, Y. & Barto, A. G. Hierarchically Organized Behavior and Its Neural Foundations: A Reinforcement Learning Perspective. *Cognition* **113,** 262–280. ISSN: 1873-7838. doi:10.1016/j.cognition.2008.08.011 (Dec. 2009).

18. Dayan, P. Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation* **5,** 613–624. ISSN: 0899-7667. doi:10.1162/neco.1993.5.4.613 (July 1993).

19. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive Representations Can Link Model-Based Reinforcement Learning to Model-Free Mechanisms. *PLOS Computational Biology* **13,** e1005768. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1005768 (Sept. 2017).

20. Momennejad, I. *et al.* The Successor Representation in Human Reinforcement Learning. *Nature Human Behaviour* **1,** 680–692. ISSN: 2397-3374. doi:10.1038/s41562-017-0180-8 (Sept. 2017).

21. Schapiro, A. C., Turk-Browne, N. B., Norman, K. A. & Botvinick, M. M. Statistical Learning of Temporal Community Structure in the Hippocampus. *Hippocampus* **26,** 3–8. ISSN: 1098-1063. doi:10.1002/hipo.22523 (2016).

22. Garvert, M. M., Dolan, R. J. & Behrens, T. E. A Map of Abstract Relational Knowledge in the Human Hippocampal–Entorhinal Cortex. *eLife* **6** (ed Davachi, L.) e17086. ISSN: 2050-084X. doi:10.7554/eLife.17086 (Apr. 2017).

23. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The Hippocampus as a Predictive Map. *Nature Neuroscience* **20,** 1643–1653. ISSN: 1546-1726. doi:10.1038/nn.4650 (Nov. 2017).

24. Kahn, A. E., Karuza, E. A., Vettel, J. M. & Bassett, D. S. Network Constraints on Learnability of Probabilistic Motor Sequences. *Nature Human Behaviour* **2,** 936–947. ISSN: 2397-3374. doi:10.1038/s41562-018-0463-8 (Dec. 2018).

25. Lynn, C. W., Kahn, A. E., Nyema, N. & Bassett, D. S. Abstract Representations of Events Arise from Mental Errors in Learning and Memory. *Nature Communications* **11,** 2313. ISSN: 2041-1723. doi:10.1038/s41467-020-15146-7 (May 2020).

26. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. *Neural Evidence for the Successor Representation in Choice Evaluation* Aug. 2021. doi:10.1101/2021.08.29.458114.

27. Ekman, M., Kusch, S. & de Lange, F. P. Successor-like Representation Guides the Prediction of Future Events in Human Visual Cortex and Hippocampus. *eLife* **12** (eds Barense, M., Baker, C. I. & Barron, H.) e78904. ISSN: 2050-084X. doi:10.7554/eLife.78904 (Feb. 2023).

28. Lieder, F. & Griffiths, T. L. Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources. *Behavioral and Brain Sciences* **43,** e1. ISSN: 0140-525X, 1469-1825. doi:10.1017/S0140525X1900061X (Jan. 2020).

29. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron* **69,** 1204–1215. ISSN: 0896-6273. doi:10.1016/j.neuron.2011.02.027 (Mar. 2011).

30. Piray, P. & Daw, N. D. Linear Reinforcement Learning: Flexible Reuse of Computation in Planning, Grid Fields, and Cognitive Control. *bioRxiv,* 856849. doi:10.1101/856849 (May 2020).

31. Piray, P. & Daw, N. D. Reconciling Flexibility and Efficiency: Medial Entorhinal Cortex Represents a Compositional Cognitive Map. *bioRxiv,* 2024–05 (2024).

32. Simon, D. & Daw, N. Environmental Statistics and the Trade-off between Model-Based and TD Learning in Humans. *Advances in Neural Information Processing Systems* **24** (2011).

33. Studer, B., Apergis-Schoute, A., Robbins, T. W. & Clark, L. What Are the Odds? The Neural Correlates of Active Choice during Gambling. *Frontiers in Neuroscience* **6.** ISSN: 1662-453X. doi:10.3389/fnins.2012.00046 (Apr. 2012).

34. Tao, R., Zhang, C., Zhao, H. & Xu, S. Active vs. Computer-Based Passive Decision-Making Leads to Discrepancies in Outcome Evaluation: Evidence from Self-Reported Emotional Experience and Brain Activity. *Cerebral Cortex* **33,** 10676–10685. ISSN: 1047-3211. doi:10.1093/cercor/bhad317 (Oct. 2023).

35. Todorov, E. Efficient Computation of Optimal Actions. *Proceedings of the National Academy of Sciences* **106,** 11478–11483. ISSN: 0027-8424, 1091-6490. doi:10.1073/pnas.0710743106 (July 2009).

36. Ho, M. K. *et al.* People Construct Simplified Mental Representations to Plan. *Nature* **606,** 129–136. ISSN: 1476-4687. doi:10.1038/s41586-022-04743-9 (June 2022).

37. Seow, T. X. F. *et al.* Model-Based Planning Deficits in Compulsivity Are Linked to Faulty Neural Representations of Task Structure. *Journal of Neuroscience* **41,** 6539–6550. ISSN: 0270-6474, 1529-2401. doi:10.1523/JNEUROSCI.0031-21.2021 (July 2021).

38. Grahek, I., Frömer, R., Prater Fahey, M. & Shenhav, A. Learning When Effort Matters: Neural Dynamics Underlying Updating and Adaptation to Changes in Performance Efficacy. *Cerebral Cortex* **33,** 2395–2411. ISSN: 1047-3211. doi:10.1093/cercor/bhac215 (Mar. 2023).

39. Zorowitz, S., Bennett, D. & jbransonbyers. *Nivlab/Nivturk: Prolific v1.2* Zenodo. June 2022. doi:10.5281/zenodo.6609218.

40. De Leeuw, J. R., Gilbert, R. A. & Luchterhandt, B. jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software* **8,** 5351. ISSN: 2475-9066. doi:10.21105/joss.05351 (May 2023).

41. Bates, D. *et al. JuliaStats/MixedModels.jl: v4.22.1* version v4.22.1. Sept. 2023. doi:10.5281/zenodo.8357196.

42. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A fresh approach to numerical computing. *SIAM Review* **59,** 65–98. doi:10.1137/141000671 (2017).

43. Huys, Q. J. M. *et al.* Disentangling the Roles of Approach, Activation and Valence in Instrumental and Pavlovian Responding. *PLoS computational biology* **7,** e1002028. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1002028 (Apr. 2011).

44. Oakes, D. Direct Calculation of the Information Matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61,** 479–482. ISSN: 1467-9868. doi:10.1111/1467-9868.00188 (1999).

1 Supplement

Full regression estimates

MB Reward Only Model:

Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	0.036	0.01	3.5	0.0004
$reward_{t-1}$	0.43	0.026	16.0	4.5e-60

SR Reward Only Model:

Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	0.023	0.01	2.3	0.021
$reward_{t-1}$	0.4	0.025	16.0	1.2e-56

SR/MB Reward Only Model:

Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	0.031	0.01	3.1	0.0019
$reward_{t-1}$	0.43	0.026	17.0	1.2e-62

Participants Reward Only Model:

Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	0.13	0.022	6.1	1.2e-9
$reward_{t-1}$	0.5	0.052	9.7	4.6e-22

Full Model SR				
Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	0.28	0.033	8.5	2.1e-17
$reward_{t-1}$	0.13	0.067	2.0	0.046
$lag1_neighborboat_reg$	-0.64	0.039	-17.0	7.7e-62
$lag2_neighborboat_reg$	-0.22	0.027	-8.1	5.3e-16
$lag3_neighborboat_reg$	-0.089	0.027	-3.3	0.00098
$lag4_neighborboat_reg$	-0.048	0.027	-1.8	0.076
$lag5_neighborboat_reg$	-0.072	0.026	-2.7	0.0067
$lag1_sameboat_reg$	0.24	0.027	8.8	1.5e-18
$lag2_sameboat_reg$	0.09	0.027	3.3	0.00088
$lag3_sameboat_reg$	-0.022	0.027	-0.81	0.42
$lag4_sameboat_reg$	0.011	0.027	0.42	0.67
$lag5_sameboat_reg$	6.2e-5	0.026	0.0024	1.0
$lag1_policy_reg$	0.012	0.03	0.41	0.68
$lag2_policy_reg$	0.0038	0.031	0.12	0.9
$lag3_policy_reg$	-0.0015	0.031	-0.047	0.96
$lag4_policy_reg$	0.089	0.031	2.9	0.0038
$lag5_policy_reg$	-0.022	0.029	-0.75	0.45
$lag1_oppislandavgboat_reg$	-1.4	0.042	-34.0	2.8e-249
$lag2_oppislandavgboat_reg$	-0.52	0.041	-13.0	5.7e-38
$lag3_oppislandavgboat_reg$	-0.22	0.04	-5.5	3.1e-8
$lag4_oppislandavgboat_reg$	-0.2	0.039	-5.1	2.8e-7
$lag5_oppislandavgboat_reg$	-0.1	0.038	-2.7	0.0076
$lag1_choice_autoreg$	0.95	0.026	37.0	6.2e-299
$lag2_choice_autoreg$	0.83	0.026	31.0	1.5e-216
$lag3_choice_autoreg$	0.13	0.027	4.7	3.0e-6
$lag4_choice_autoreg$	0.34	0.027	13.0	5.1e-38
$lag5_choice_autoreg$	0.1	0.026	3.8	0.00013
$reward_{t-1}$ & $lag1_neighborboat_reg$	-0.088	0.055	-1.6	0.11
$reward_{t-1}$ & $lag2_neighborboat_reg$	-0.0027	0.054	-0.05	0.96
$reward_{t-1}$ & $lag3_neighborboat_reg$	-0.049	0.054	-0.91	0.36
$reward_{t-1}$ & $lag4_neighborboat_reg$	-0.11	0.054	-2.0	0.047
$reward_{t-1}$ & $lag5_neighborboat_reg$	-0.012	0.053	-0.23	0.82
$reward_{t-1}$ & $lag1_sameboat_reg$	0.19	0.054	3.5	0.00044
$reward_{t-1}$ & $lag2_sameboat_reg$	-0.0096	0.054	-0.18	0.86
$reward_{t-1}$ & $lag3_sameboat_reg$	-0.026	0.054	-0.49	0.62
$reward_{t-1}$ & $lag4_sameboat_reg$	0.079	0.053	1.5	0.14
$reward_{t-1}$ & $lag5_sameboat_reg$	-0.099	0.052	-1.9	0.059
$reward_{t-1}$ & $lag1_policy_reg$	0.7	0.08	8.7	3.1e-18
$reward_{t-1}$ & $lag2_policy_reg$	0.2	0.063	3.1	0.0018
$reward_{t-1}$ & $lag3_policy_reg$	0.14	0.062	2.2	0.025
$reward_{t-1}$ & $lag4_policy_reg$	0.16	0.062	2.6	0.011
$reward_{t-1}$ & $lag5_policy_reg$	0.15	0.058	2.5	0.013

Full Model MB				
Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	0.19	0.031	5.9	2.7e-9
$reward_{t-1}$	-0.43	0.056	-7.6	2.1e-14
$lag1_neighborboat_reg$	-0.68	0.04	-17.0	3.3e-64
$lag2_neighborboat_reg$	-0.22	0.027	-8.3	1.4e-16
$lag3_neighborboat_reg$	-0.15	0.026	-5.6	2.0e-8
$lag4_neighborboat_reg$	-0.027	0.026	-1.0	0.3
$lag5_neighborboat_reg$	-0.026	0.026	-1.0	0.31
$lag1_sameboat_reg$	0.38	0.026	14.0	1.1e-46
$lag2_sameboat_reg$	0.11	0.027	4.2	2.2e-5
$lag3_sameboat_reg$	0.023	0.026	0.88	0.38
$lag4_sameboat_reg$	0.0011	0.026	0.043	0.97
$lag5_sameboat_reg$	-0.0027	0.026	-0.1	0.92
$lag1_policy_reg$	0.0086	0.032	0.27	0.79
$lag2_policy_reg$	-0.0073	0.032	-0.23	0.82
$lag3_policy_reg$	-0.061	0.032	-1.9	0.058
$lag4_policy_reg$	0.03	0.031	0.96	0.34
$lag5_policy_reg$	0.023	0.029	0.78	0.44
$lag1_oppislandavgboat_reg$	-1.5	0.039	-39.0	0.0
$lag2_oppislandavgboat_reg$	-0.55	0.039	-14.0	5.8e-45
$lag3_oppislandavgboat_reg$	-0.2	0.039	-5.2	1.8e-7
$lag4_oppislandavgboat_reg$	-0.19	0.038	-5.0	4.7e-7
$lag5_oppislandavgboat_reg$	-0.031	0.037	-0.84	0.4
$lag1_choice_autoreg$	0.8	0.025	32.0	4.9e-228
$lag2_choice_autoreg$	0.44	0.025	17.0	9.0e-68
$lag3_choice_autoreg$	0.065	0.026	2.5	0.011
$lag4_choice_autoreg$	0.11	0.025	4.3	1.9e-5
$lag5_choice_autoreg$	0.043	0.025	1.7	0.084
$reward_{t-1}$ & $lag1_neighborboat_reg$	0.8	0.066	12.0	9.4e-34
$reward_{t-1}$ & $lag2_neighborboat_reg$	0.5	0.053	9.6	1.1e-21
$reward_{t-1}$ & $lag3_neighborboat_reg$	0.13	0.053	2.4	0.016
$reward_{t-1}$ & $lag4_neighborboat_reg$	0.011	0.052	0.21	0.83
$reward_{t-1}$ & $lag5_neighborboat_reg$	-0.0038	0.052	-0.072	0.94
$reward_{t-1}$ & $lag1_sameboat_reg$	0.43	0.052	8.3	1.1e-16
$reward_{t-1}$ & $lag2_sameboat_reg$	0.11	0.053	2.2	0.03
$reward_{t-1}$ & $lag3_sameboat_reg$	0.16	0.053	3.0	0.0024
$reward_{t-1}$ & $lag4_sameboat_reg$	0.015	0.052	0.28	0.78
$reward_{t-1}$ & $lag5_sameboat_reg$	0.0063	0.052	0.12	0.9
$reward_{t-1}$ & $lag1_policy_reg$	-0.0028	0.067	-0.042	0.97
$reward_{t-1}$ & $lag2_policy_reg$	0.02	0.064	0.32	0.75
$reward_{t-1}$ & $lag3_policy_reg$	-0.067	0.065	-1.0	0.3
$reward_{t-1}$ & $lag4_policy_reg$	0.025	0.063	0.4	0.69
$reward_{t-1}$ & $lag5_policy_reg$	0.026	0.058	0.45	0.65

SR/MB Hybrid Model				
Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	0.22	0.032	7.1	1.8e-12
$reward_{t-1}$	-0.18	0.064	-2.8	0.0055
$lag1_neighborboat_reg$	-0.72	0.037	-19.0	6.4e-82
$lag2_neighborboat_reg$	-0.23	0.027	-8.6	5.7e-18
$lag3_neighborboat_reg$	-0.14	0.027	-5.4	7.9e-8
$lag4_neighborboat_reg$	-0.064	0.026	-2.4	0.015
$lag5_neighborboat_reg$	-0.029	0.026	-1.1	0.27
$lag1_sameboat_reg$	0.31	0.027	12.0	1.9e-32
$lag2_sameboat_reg$	0.1	0.027	3.9	9.4e-5
$lag3_sameboat_reg$	0.0014	0.027	0.052	0.96
$lag4_sameboat_reg$	0.013	0.026	0.5	0.62
$lag5_sameboat_reg$	0.034	0.026	1.3	0.19
$lag1_policy_reg$	0.031	0.03	1.0	0.3
$lag2_policy_reg$	0.022	0.03	0.72	0.47
$lag3_policy_reg$	0.047	0.031	1.6	0.12
$lag4_policy_reg$	0.0093	0.03	0.31	0.76
$lag5_policy_reg$	-0.015	0.028	-0.53	0.59
$lag1_oppislandavgboat_reg$	-1.5	0.041	-37.0	6.4e-292
$lag2_oppislandavgboat_reg$	-0.52	0.04	-13.0	4.4e-40
$lag3_oppislandavgboat_reg$	-0.25	0.039	-6.5	9.9e-11
$lag4_oppislandavgboat_reg$	-0.16	0.038	-4.2	3.2e-5
$lag5_oppislandavgboat_reg$	-0.078	0.038	-2.0	0.04
$lag1_choice_autoreg$	0.92	0.025	37.0	5.3e-295
$lag2_choice_autoreg$	0.55	0.026	21.0	7.8e-99
$lag3_choice_autoreg$	0.18	0.026	6.7	2.3e-11
$lag4_choice_autoreg$	0.11	0.026	4.2	2.5e-5
$lag5_choice_autoreg$	0.11	0.026	4.4	1.2e-5
$reward_{t-1}$ & $lag1_neighborboat_reg$	0.34	0.064	5.4	7.2e-8
$reward_{t-1}$ & $lag2_neighborboat_reg$	0.12	0.053	2.2	0.027
$reward_{t-1}$ & $lag3_neighborboat_reg$	0.076	0.053	1.4	0.15
$reward_{t-1}$ & $lag4_neighborboat_reg$	0.12	0.053	2.3	0.024
$reward_{t-1}$ & $lag5_neighborboat_reg$	0.069	0.052	1.3	0.19
$reward_{t-1}$ & $lag1_sameboat_reg$	0.32	0.053	6.0	1.8e-9
$reward_{t-1}$ & $lag2_sameboat_reg$	0.075	0.053	1.4	0.16
$reward_{t-1}$ & $lag3_sameboat_reg$	-0.0044	0.053	-0.083	0.93
$reward_{t-1}$ & $lag4_sameboat_reg$	-0.0099	0.053	-0.19	0.85
$reward_{t-1}$ & $lag5_sameboat_reg$	-0.11	0.052	-2.1	0.037
$reward_{t-1}$ & $lag1_policy_reg$	0.54	0.082	6.6	4.9e-11
$reward_{t-1}$ & $lag2_policy_reg$	0.14	0.062	2.3	0.02
$reward_{t-1}$ & $lag3_policy_reg$	0.00041	0.062	0.0066	0.99
$reward_{t-1}$ & $lag4_policy_reg$	0.13	0.06	2.2	0.028
$reward_{t-1}$ & $lag5_policy_reg$	0.021	0.057	0.38	0.71

Participants				
Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	0.21	0.052	4.1	3.9e-5
$reward_{t-1}$	0.00019	0.093	0.002	1.0
$lag1_neighborboat_reg$	-0.69	0.053	-13.0	8.2e-39
$lag2_neighborboat_reg$	-0.14	0.041	-3.3	0.00089
$lag3_neighborboat_reg$	-0.12	0.04	-2.9	0.0035
$lag4_neighborboat_reg$	-0.034	0.04	-0.86	0.39
$lag5_neighborboat_reg$	-0.12	0.04	-2.9	0.0035
$lag1_sameboat_reg$	0.48	0.04	12.0	2.9e-32
$lag2_sameboat_reg$	0.17	0.04	4.2	2.3e-5
$lag3_sameboat_reg$	0.05	0.04	1.2	0.21
$lag4_sameboat_reg$	0.1	0.04	2.5	0.012
$lag5_sameboat_reg$	0.022	0.04	0.54	0.59
$lag1_policy_reg$	-0.038	0.044	-0.85	0.4
$lag2_policy_reg$	-0.058	0.044	-1.3	0.19
$lag3_policy_reg$	0.03	0.046	0.65	0.52
$lag4_policy_reg$	0.0094	0.046	0.2	0.84
$lag5_policy_reg$	0.046	0.044	1.1	0.29
$lag1_oppislandavgboat_reg$	-1.6	0.061	-25.0	5.9e-142
$lag2_oppislandavgboat_reg$	-0.32	0.06	-5.4	6.2e-8
$lag3_oppislandavgboat_reg$	-0.13	0.058	-2.2	0.027
$lag4_oppislandavgboat_reg$	-0.13	0.058	-2.2	0.027
$lag5_oppislandavgboat_reg$	-0.12	0.058	-2.0	0.045
$lag1_choice_autoreg$	1.3	0.04	32.0	1.5e-227
$lag2_choice_autoreg$	0.95	0.042	23.0	1.9e-116
$lag3_choice_autoreg$	0.06	0.043	1.4	0.16
$lag4_choice_autoreg$	0.49	0.042	12.0	1.0e-30
$lag5_choice_autoreg$	0.057	0.041	1.4	0.17
$reward_{t-1}$ & $lag1_neighborboat_reg$	0.74	0.11	6.8	1.3e-11
$reward_{t-1}$ & $lag2_neighborboat_reg$	0.13	0.081	1.6	0.11
$reward_{t-1}$ & $lag3_neighborboat_reg$	0.018	0.08	0.22	0.83
$reward_{t-1}$ & $lag4_neighborboat_reg$	-0.067	0.08	-0.85	0.4
$reward_{t-1}$ & $lag5_neighborboat_reg$	-0.053	0.079	-0.66	0.51
$reward_{t-1}$ & $lag1_sameboat_reg$	0.12	0.08	1.5	0.15
$reward_{t-1}$ & $lag2_sameboat_reg$	-0.016	0.08	-0.2	0.84
$reward_{t-1}$ & $lag3_sameboat_reg$	-0.042	0.08	-0.52	0.6
$reward_{t-1}$ & $lag4_sameboat_reg$	0.053	0.08	0.67	0.5
$reward_{t-1}$ & $lag5_sameboat_reg$	0.021	0.079	0.26	0.79
$reward_{t-1}$ & $lag1_policy_reg$	0.38	0.091	4.2	2.5e-5
$reward_{t-1}$ & $lag2_policy_reg$	0.12	0.089	1.3	0.19
$reward_{t-1}$ & $lag3_policy_reg$	0.26	0.092	2.8	0.005
$reward_{t-1}$ & $lag4_policy_reg$	-0.11	0.091	-1.2	0.22
$reward_{t-1}$ & $lag5_policy_reg$	-0.18	0.087	-2.0	0.043

Model recovery when SR behavior is absent

We wished to verify that no aspect of our experimental design biased model recovery towards estimating SR-like behavior if not present. To explore this possibility, we simulated a set of agents whose parameters were drawn from the estimates of a variation of non-blockwise model which did not model any SR behavior, but included a single β_{MB} and β_{TD}. We simulated 100 agents, each as a unique draw from the recovered mean and variance structure of the human data, and then sought to recover β_{MB}, β_{TD}, and β_{SR}. In Fig. 5, we plot the distributions of recovered group-level mean for β_{SR} ($\mu = 0.01, \sigma = 0.073$), as well as the estimate from our recovered participant data ($\mu = 0.866$).

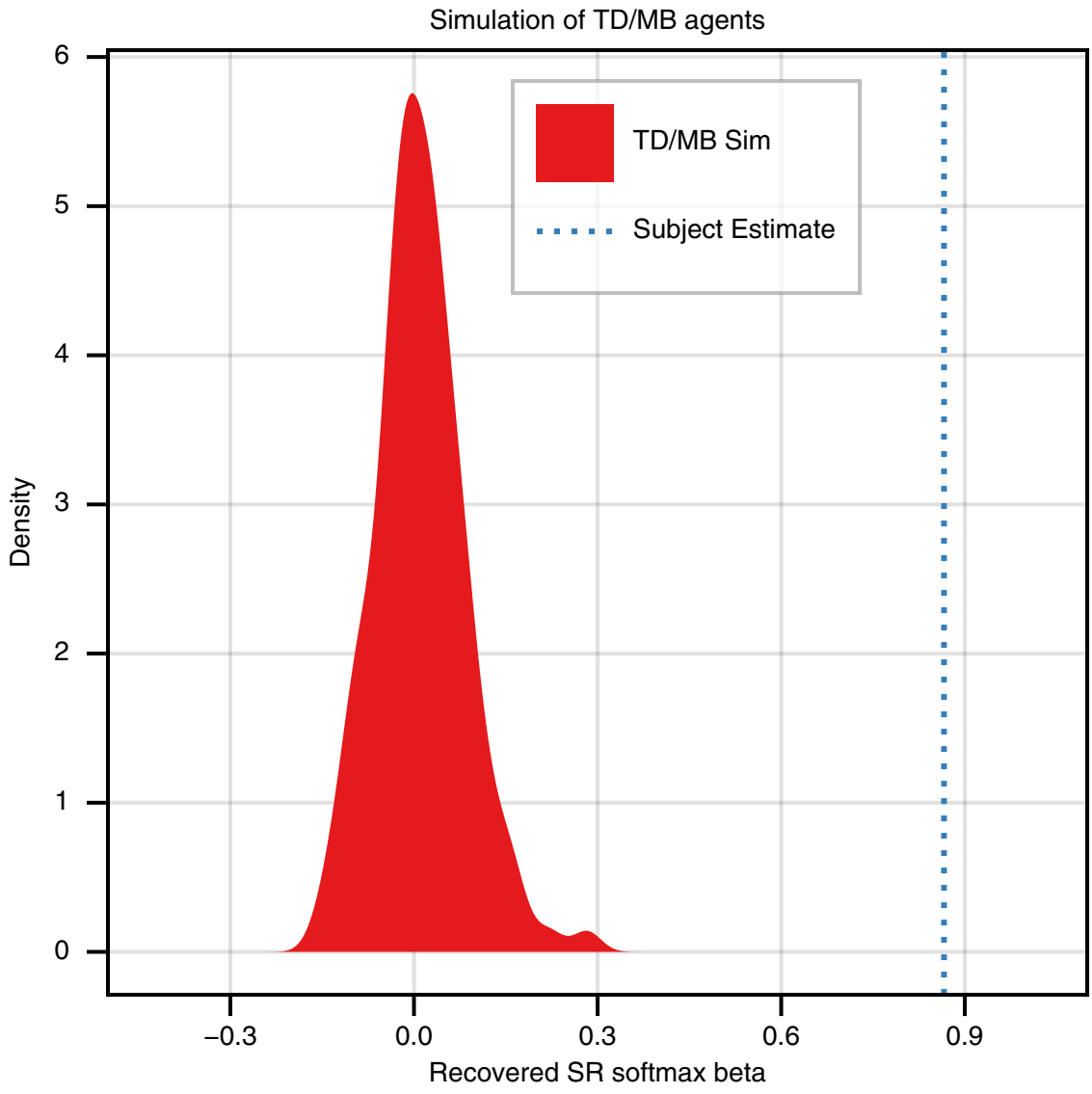

Figure 5: **Recovery of null model simulations for SR behavior**. Distribution of estimated values of SR softmax temperature for simulated pools of TD/MB agents. The dotted line represents the value of the SR softmax temperature estimated from human participants.

Blockwise trial structure does not lead to recovery of distinct blockwise softmax parameters

A potential concern with our experimental manipulation is that the discrepancy in changed reward between our two trial types (incongruent block changes lead to a larger loss in reward when following the existing policy) may lead to a model-free shift that masquerades as a change in reliance on prior policy, and thus the MB/SR balance between blocks. To explore this possibility, we simulated a set of agents whose parameters were drawn from the estimates of our non-blockwise model: a single β_{SR}, β_{MB}, and β_{TD}. We simulated 100 agents, each as a unique draw from the recovered mean and variance structure of the human data, and then sought to recover blockwise parameters using our weighted model. We did not find a significant difference in $w_{\text{SR}}+$ and $w_{\text{SR}}-$ (Fig. 6): we found an expected difference of 0.269, 95% CI = +/- 0.57, $t_1044 = 0.92$, $p < 0.36$.

Linear RL parameter estimates for MB and SR simulations

Linear RL converges to MB behavior as $\lambda \to 0$, and to SR behavior as $\lambda \to \infty$. However, in practice, we expect to observe neither of these extremes when fitting $\log(\lambda)$ to human behavior. What values are representative of MB and SR behavior? To answer this question, we simulated agents whose parameters were drawn from the estimates of our non-blockwise-model when assuming either purely MB or purely SR behavior (that is, estimating only β_{MB} in one case and only β_{SR} in the other). We simulated 100 agents for each case, each as a unique draw from the recovered mean and variance structure of the human data, and then sought to recover a group-level estimate of $\log(\lambda)$ for each scenario. In Fig. 7, we plot the distributions of recovered group-level means for $\log(\lambda)$ (MB: $\mu = -2.347, \sigma = 0.359$, SR: $\mu = 2.742, \sigma = 0.661$, as well as the estimate from our recovered participant data ($\mu = -0.802$).

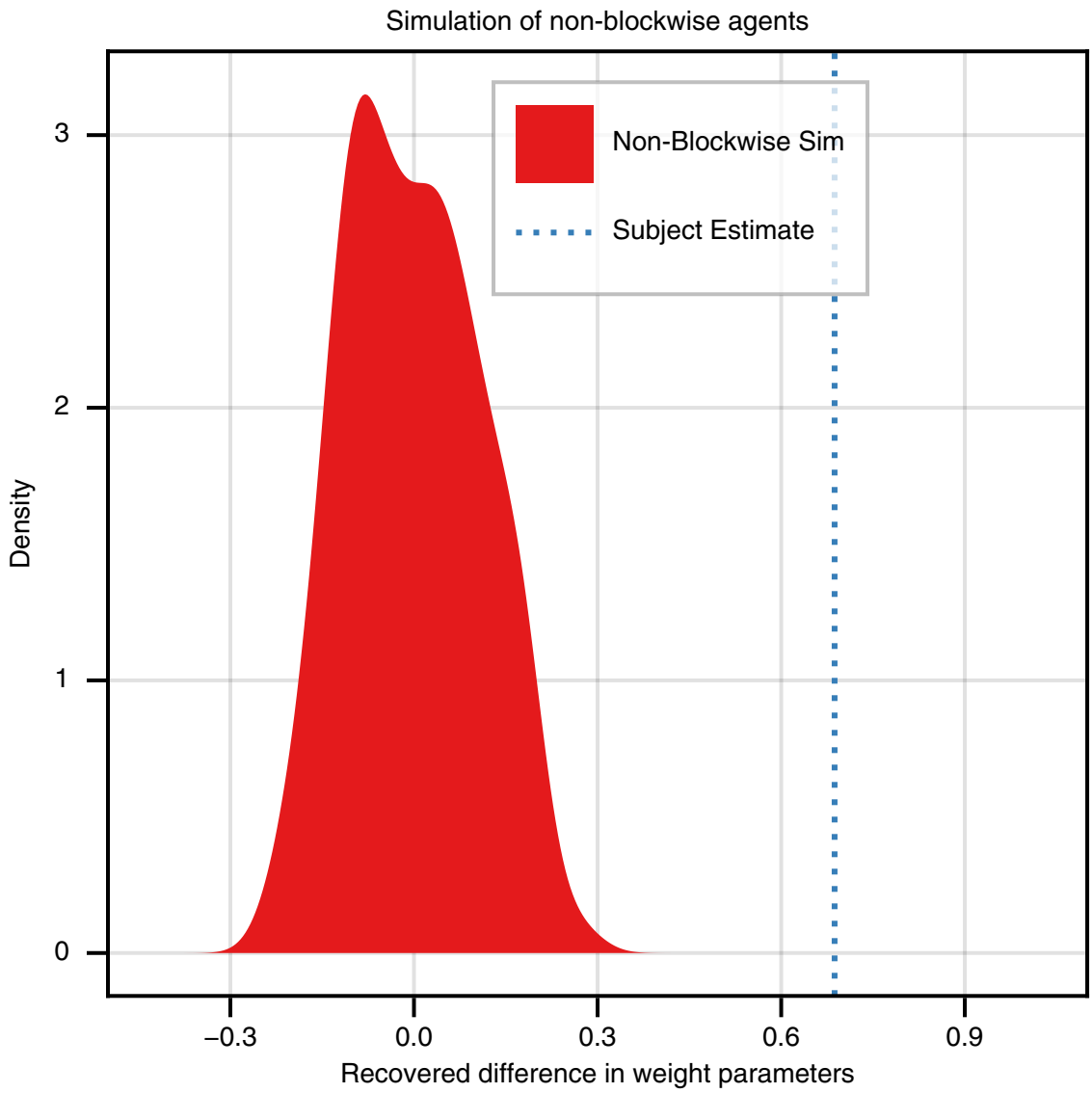

Figure 6: **Recovery of null model simulations for blockwise behavior**. Distribution of estimated values of the differences in fit weight parameters for simulated pools of MB/SR/TD(1) agents without blockwise behavior. The dotted line represents the value of the weight difference estimated from human participants.

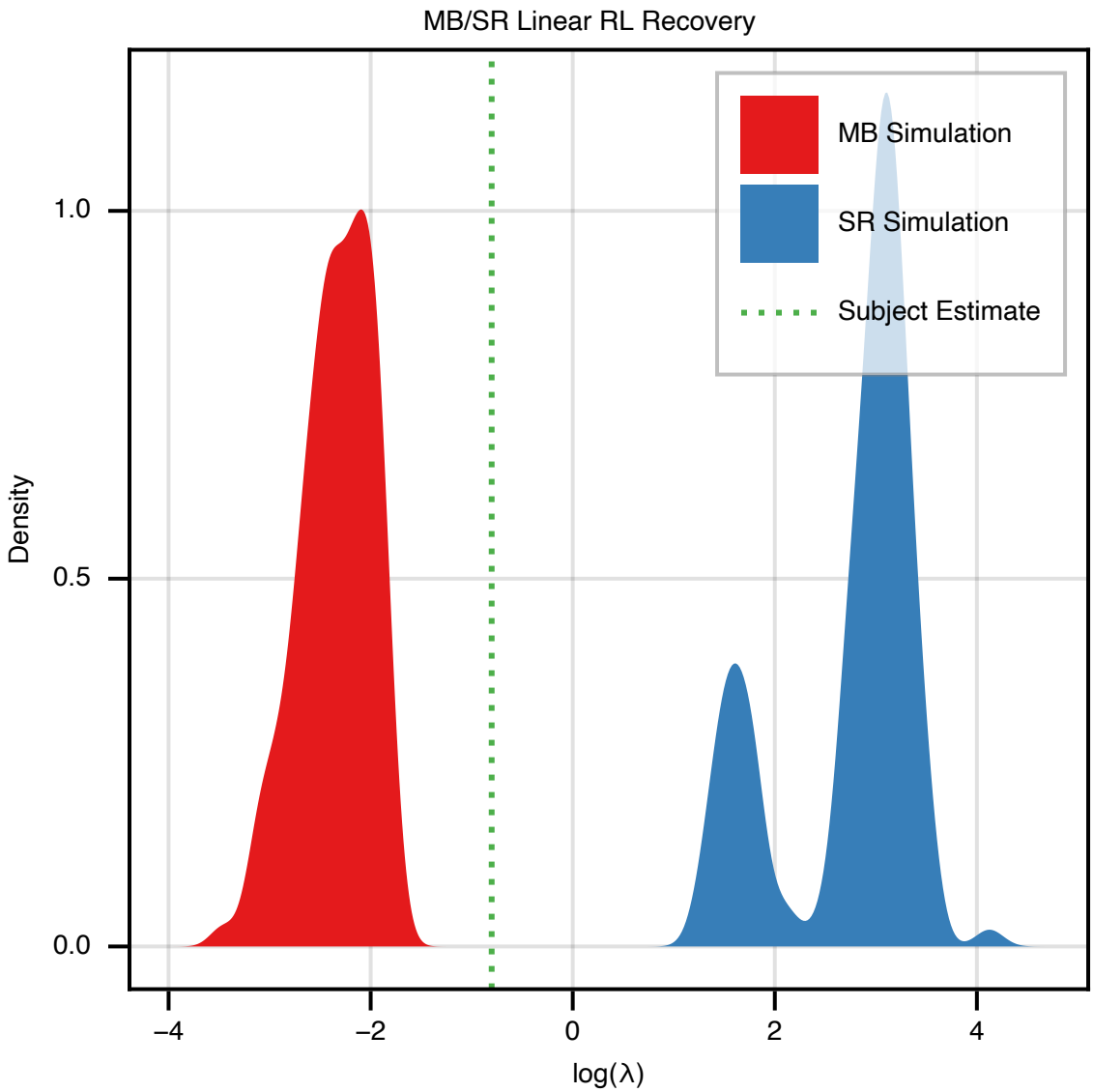

Figure 7: **Linear RL parameter estimates**. Distribution of estimated values of $\log(\lambda)$ for simulated pools of either MB (red) or SR (blue) agents. The dotted line represents the value of $\log(\lambda)$ estimated from human participants.