Humans rationally balance mental simulation and temporally-abstract heuristics

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

November 28, 2023

1 Abstract

When faced with a multi-step decision problem, humans and animals must balance flexible and accurate decision making with computational complexity. One prominent approach, the Successor Representation (SR), takes advantage of temporal abstraction of future states: by learning to predict long-run future trajectory independently of rewards, the brain can avoid the costs of iterative, multi-step model-based mental simulation, while retaining some ability to cheaply replan when goals change. Human behavior shows signatures of such temporal abstraction, but the characterization of the strategy of individuals, as well as whether people dynamically adapt their reliance on such abstractions in the face of environmental statistics, e.g. the predictability of long-run states, 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 structure within-subject, we observe evidence of resource-rational reliance on the SR, which decreases when the ability to use prior experience to build valid temporal abstractions decreases. 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 predictibility of their environment.

Keywords: successor representation; model-based; reinforcement learning; decision making; planning

2 Introduction

A hallmark of human and animal planning is its flexibility: the ability to effectively plan and replan in the face of novel or changing circumstances and outcomes. 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. A long-considered solution to this problem is for the brain to simulate outcomes through use a cognitive map or internal model (Tolman, 1948). Reinforcement learning (RL) theories formalize this idea in terms of "model-based" (MB) algorithms, which employ a learned model of the short-term consequences of actions in order to iteratively 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 (Sutton, 1988) 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 (Daw, Niv, & Dayan, 2005; Keramati, Dezfouli, & Piray, 2011; Kool, Gershman, & Cushman, 2017). This approach suggests a formal, resource-rational account of both healthy and pathological habits (Everitt & Robbins, 2005; Gillan, Kosinski, Whelan, Phelps, & Daw, 2016).

The successes of this framework as well as the practical limitations of both extreme algorithms

have motivated the study of more nimble heuristics 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 (Botvinick, Niv, & Barto, 2009). Prominent among these is the Successor Representation (SR) (Dayan, 1993), a planning algorithm that models future states (instead of future rewards) aggregated over multiple timesteps. This allows quickly recomputing net payoffs (e.g., if rewards change) without step-by-step simulation (Russek, Momennejad, Botvinick, Gershman, & Daw, 2017). 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 (Schapiro, Turk-Browne, Norman, & Botvinick, 2016; Garvert, Dolan, & Behrens, 2017; Stachenfeld, Botvinick, & Gershman, 2017; Momennejad et al., 2017; Lynn, Kahn, Nyema, & Bassett, 2020; Russek, Momennejad, Botvinick, Gershman, & Daw, 2021; Ekman, Kusch, & de Lange, 2023). However, it remains unclear whether the concept of resource-rational tradeoffs (e.g. balancing 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 distinguish SR from MB (Momennejad et al., 2017). This makes it difficult to measure between-individual or -condition variation in SR usage, as well as tracking many 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 (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). 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 then characterize individual subjects' relative reliance on either approach using Linear Reinforcement Learning (Linear RL) (Piray & Daw, 2020), a biologically-plausible planning algorithm that generalizes MB and SR, incorporating them as extreme cases in a spectrum of behaviors ranging from fully relying on the policy prior (in our task, identical to the SR) to fully MB tree search (Fig. 1a, center), with the range between them captured by a single parameter, λ.

The ability to characterize graded SR vs. MB usage within individuals allows us to investigate whether people rationally adjust their reliance on the SR in response to circumstances such as environmental predictability. Whereas the utility of MF learning versus MB planning depends on the stability of action value estimates (Simon & Daw, 2011), 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 λ while manipulating these temporal statistics, and found that people rationally adjusted λ to match the accuracy of the SR. This suggests that humans can arbitrate between temporally-abstracted heuristics and MB planning in response to their respective utility, adding to a growing literature on rational balance between planning mechanisms.

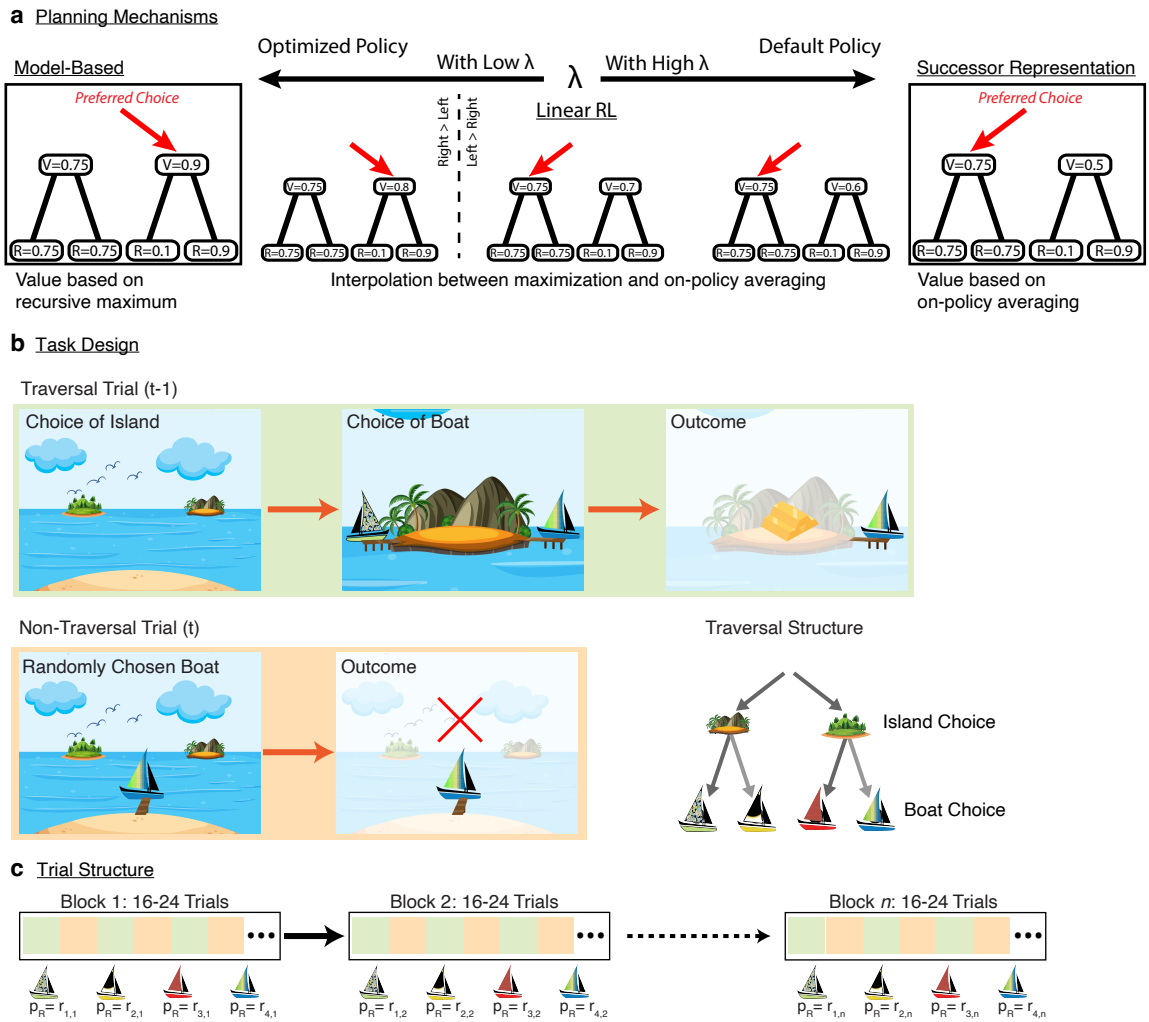

Figure 1: **Planning Model and Task Structure**. (**a**) A MB agent (left) evaluates choices via recursive maximization, while an SR agent (right) evaluates choices via on-policy averaging. Linear RL (center) interpolates between these two extremes via a single parameter, λ. (**b**) Trials alternate 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. (**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.

3 Results

3.1 Behavior

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 (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. 1b, top), whereas non-traversal trials involved no decision making, with one of the boats visited at random (Fig. 1b, 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.

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.

3.2 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. 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.

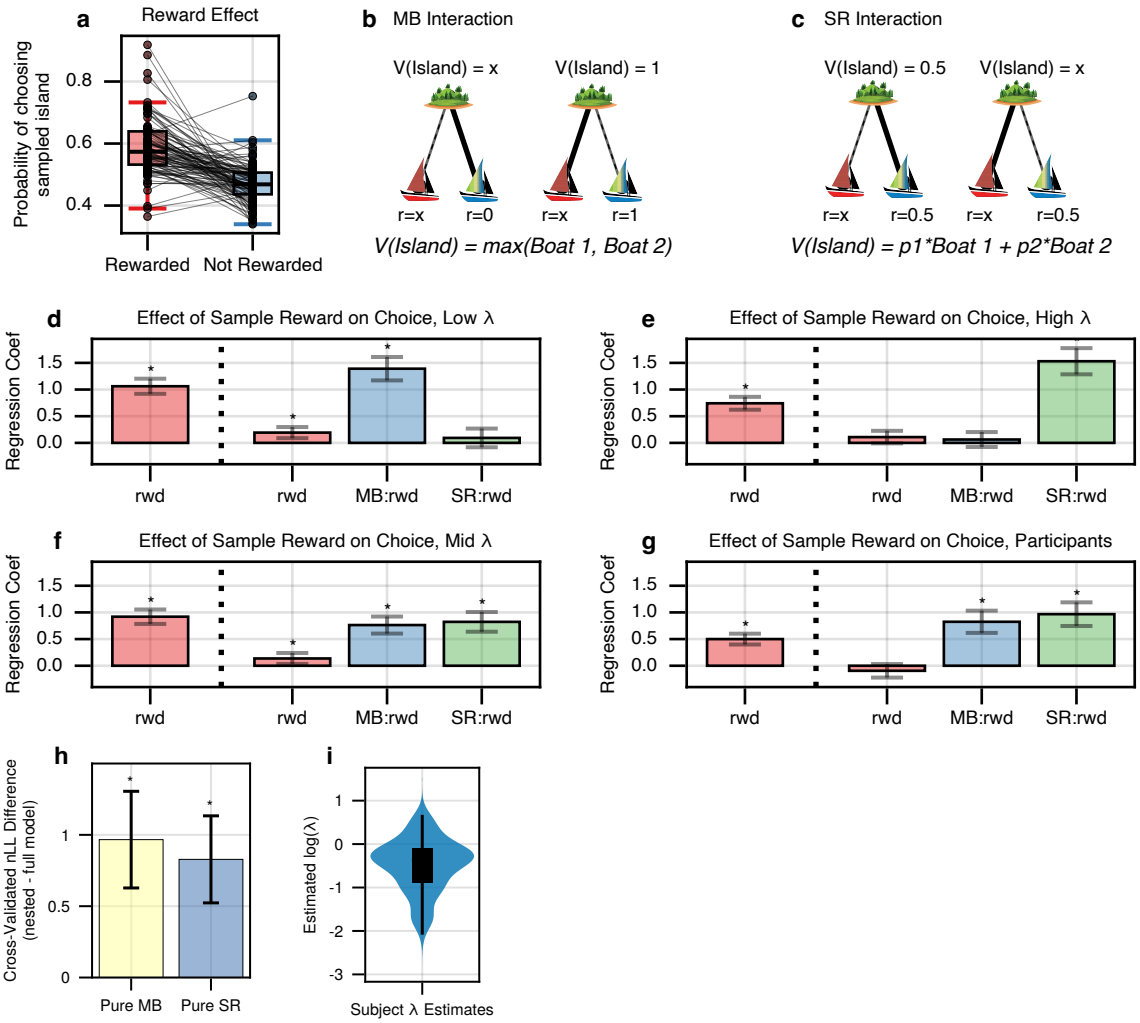

Figure 2: **Distinct Signatures of MB and SR**. (**a**) Reward Sensitivity: Participants are significantly more likely to choose the island associated with a sampled boat when the boat was rewarded, behavior which cannot be explained by MF planning mechanisms. (**b**) MB signature: A MB agent (left) maximizes over boat values at an island, so it is only sensitive to a boat's reward (vs non-reward) if the neighboring boat has low value (recently unrewarded). (**c**) SR signature: A SR agent (center) 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. Low-λ learners show a strong interaction with the neighboring boat, whereas high-λ learners show a strong interaction with prior choice at the sampled island. Middle-λ agents, as well as humans, exhibit both interaction effects. (**h**) Gain or loss of model fit when λ is constrained to either be MB or SR-like. (**i**) Estimated per-subject fits of $log(\lambda)$.

3.3 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 $Boat1$, $Boat2$ etc.) to construct (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 $Boat1$'s reward should have less effect on the MB-estimated value $V(island) = max(Boat1, Boat2)$ if the neighboring $Boat2$ 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 pre-existing ("default") choice policy at the island: thus, $Boat1's$ updated contribution to the SR-computed value $V(island) = p1 * Boat1 + p2 * Boat2$ is weighted by $p1$, 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 $Boat1$'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 ($Boat2$ and $p1$, 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 $Boat2$ by the reward (0 or 1) received the last time that boat was encountered; and we similarly approximate the default choice policy $p1$ by a binary indicator (0 or 1) for whether $Boat1$ was chosen the last time its island was encountered.

Fig. 2d-f demonstrates these effects on simulated choice data for 200 agents simulated with Linear RL using different values of the parameter λ: either favoring MB planning (low λ, Fig. 2d), favoring SR (high λ, Fig. 2e) or an intermediate mixture of MB and SR (mid λ, Fig. 2f). 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 default policy (SR), and 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 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 ($\beta = -0.0948$, 95% CI = +/- 0.127, $p < 0.143$), whereas both interactions with reward show significant positive effects (MB effect: $\beta = 0.824$, 95% CI = +/- 0.208, $p < 8.58e-15$,

SR effect: $\beta = 0.967$, 95% CI = +/- 0.221, $p < 1.14e-17$). Notably, this mixture of MB and SR behavior is in-line with previous work such as (Momennejad et al., 2017), which also suggested (using a much different task and block-wise 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.

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 full model that estimated island values via Linear RL (Piray & Daw, 2020), a reinforcement learning strategy which allows for continuous interpolation between model-based and policy-based extremes via a parameter λ, where the limit of low λ corresponds to a MB estimate, and increasing λ shifts this toward a default policy (which here we take as the recently performed one), converging in the limit of high λ to the on-policy SR (Fig. 1a, center). We compared this model's fit to choice behavior (in which λ could freely vary) to two nested models (fixed extreme λ) that capture pure MB or SR behavior (Fig. 2h). The full model significantly outperformed both fixed MB (paired t-test across subjects; est. $= -0.966$, 95% CI = +/- $0.673, t_{99} = -2.85, p < 0.006$) and SR (paired t-test; est. $= -0.828$, 95% CI = +/- $0.605, t_{99} = -2.715, p < 0.008$), using leave-one-out cross validation at the subject level (see Methods). Individual subject fits are shown in Fig. 2i. Consistent with the model-agnostic regressions, this result indicates that human behavior exhibits distinct characteristics of both MB and SR planning, and that choice behavior is best-fit by a model that allows for a mixture of the two.

3.4 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 (Daw et al., 2005; Keramati et al., 2011; Kool et al., 2017), 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 are 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, top). Second, rewards might undergo an *incongruent change, 6/21 block changes*, which changed the optimal boat choice at each island (Fig. 3a, bottom). (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.)

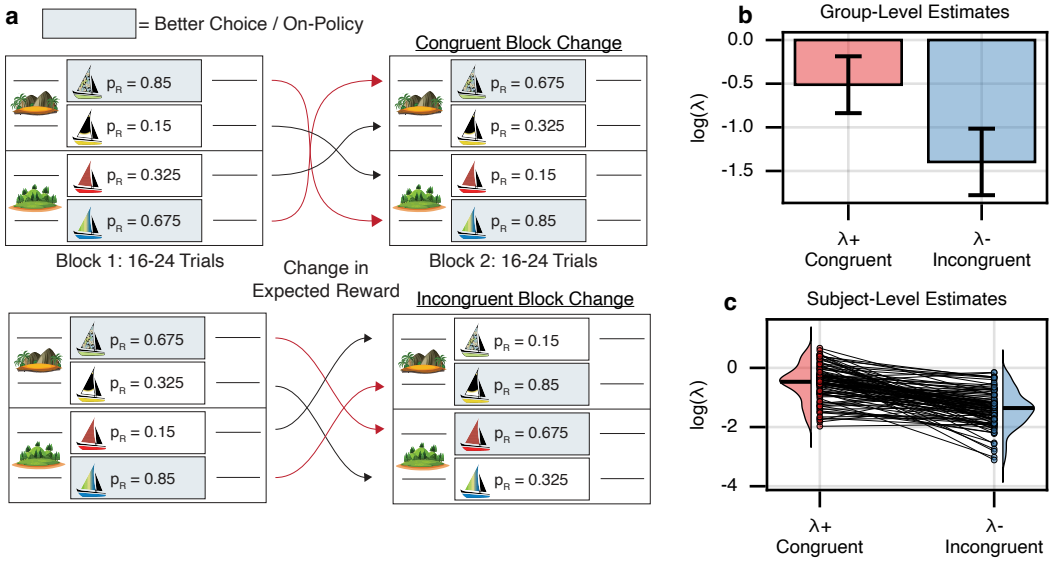

Figure 3: λ **systematically varies across block reward structure**. (**a**) Top: Congruent block change. After 16-24 trials, reward probabilities swapped such that the optimal choice of island flipped, but the optimal boat choice conditioned on island choice remained unchanged. Crucially, evaluating the new rewards via the policy induced by previous rewards still provided the optimal choice of island. Bottom: 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 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. (**c**) Individual subject fits. Each line is one subject.

To estimate whether subjects' balance of MB and SR planning systematically varied across blocks, we fit subject behavior with a model that estimated island values via Linear RL (Piray & Daw, 2020), with separate λ parameters for each block determined by which type of reversal began it. 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 high value of λ, indicating SR usage. If, instead, new reward values led to incorrect choice of islands under the prior policy, participants would switch to a lower value of λ, indicating more effort spent on explicit MB planning. To test for blockwise changes in estimation, we modeled each subject as estimating value with one of two λ parameters: $\lambda+$ was used to estimate value on consistent blocks, and $\lambda-$ was used to estimate value on inconsistent blocks, and we allowed both to freely vary. Notably, the model was free to

fit $\lambda+$ as either larger or smaller than $\lambda-$; the only difference was on which set of blocks each was used for estimation.

Indeed, we found using a hierarchical model that the group-level mean $\lambda+$ (the value following a congruent reward change) was significantly larger than $\lambda-$ (the value following an incongruent reward change): $\log(\lambda+) = -0.512$, 95% CI = +/- 0.326, $\log(\lambda-) = -1.396$, 95% CI= +/- 0.381, and the difference between the two was estimated as $0.884, 95\%CI = +/-0.478$, two-sided $t_{728} = 3.629, p < 0.0004$, (Fig. 3b). Accordingly, on blocks following congruent reward changes, subjects exhibited increased reliance on temporally-abstracted planning, whereas following incongruent reward changes, subjects turned more to MB evaluation. We observed this effect at the subject level as well: 86/100 participants were individually best explained via $\log(\lambda+) > \log(\lambda-)$ (Fig. 3c). 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.

4 Discussion

When do we spend time and mental resources explicitly simulating the future, and when do we rely on simplifying heuristics? Heuristics inherently trade off planning complexity against accuracy, and the usefulness of any given heuristic for planning depends on the environment an agent is operating within (Daw et al., 2005). For example, MF learning (or more generally, any method that caches an expectation for the long-run outcome of an action) is advantageous when such long-run values change slowly enough for a cached estimate to accurately reflect distant reward outcomes (Simon & Daw, 2011). It has been hypothesized that the brain trades off such strategies judiciously according to circumstance, and (in the case of MB vs MF arbitration) this process explains phenomena of both healthy automaticity and pathological compulsion (Daw et al., 2005; Gillan et al., 2016; Seow et al., 2021).

Here we extended this program to a nimbler (but non-exclusive) class of heuristics that have seen recent attention in decision neuroscience: those, including the SR, that simplify planning not by ignoring state dynamics altogether but instead by collapsing the iterative simulation of dynamics by using temporal abstraction over future state encounters: in effect, a simplified world model (Dayan, 1993; Piray & Daw, 2020; Russek et al., 2017). These 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 (Momennejad et al., 2017). Using these methods, we confirmed previous suggestions that people use such temporally abstracted heuristics (the SR) alongside 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.

4.1 Trial-by-Trial Measure 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. To test for this limitation, previous approaches have relied on one-shot train-retrain-probe designs ("policy revaluation" retraining), contrasting one-shot change scenarios where a 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 (Momennejad et al., 2017). 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 prior 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 timecourse of the experiment. Extending previous approaches that used similar logic and designs to distinguish MB from MF updating (Daw et al., 2011), 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 tradeoffs between multiple planning systems, and, via better characterization of individual differences in these behaviors, to investigate whether usage of SR systematically varies in psychiatric conditions, as well as over the course of normal human development.

4.2 Tradeoffs in Linear RL

In the current work, we use Linear RL (Piray & Daw, 2020) primarily as a convenient *descriptive* model for quantifying behavior along a spectrum between MB and SR planning. Of course, other mechanisms for mixing these effects are possible; for instance, we see similar qualitative results fitting behavior to weighted mixtures of separate MB and SR agents, analogous to mixture-of-agents models used for MB vs. MF behavior (Daw et al., 2011). Equally, it would be possible to blend both effects using discrete steps of MB replay (Sutton, 1991) to gradually update an SR ("SR-Dyna"; Russek et al., 2021) as again has been imagined as a different way to blend MB and MF (Mattar & Daw, 2018).

Thus, the present results do not speak directly to the mechanism by which MB- and SR-like effects are combined in the brain, or specifically support Linear RL over alternatives. That being said, linear RL provides an appealing, biologically realizable framework for implementing both computations and their tuneable tradeoff. Importantly, Linear RL makes explicit the tradeoffs between MB and SR planning, since increasing λ effectively scales the precision of reward representations inside the computations, and might be identified concretely with something like the energetic costs of using fewer or more spikes to represent decision variables (Piray & Daw, 2020).

4.3 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 (Keramati et al., 2011; Kool et al., 2017) between fast heuristics 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 (Simon & Daw, 2011). 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 (Daw et al., 2005), or on estimates of sub-optimality of reward obtained under the current policy (Kool et al., 2017). 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.

5 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 (Zorowitz, Bennett, & jbransonbyers, 2022) and jsPsych (de Leeuw, Gilbert, & Luchterhandt, 2023) 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 Manipulation**).

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. 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 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 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, 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 policy-consistent swap, a policy optimized for the previous reward structure will still correctly identify which island is better under the new rewards, but for the policy-inconsistent and within-island swaps, 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.

5.1 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 (Bates et al., 2023) in *Julia* 1.9.3 (Bezanson, Edelman, Karpinski, & Shah, 2017) with a Logit link function and the following formula::

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

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, including all two-way interaction terms and full random effects on all model coefficients:

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

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. All regressors are relative to the boat that was just sampled and the island on which that boat is found, and $sibNoRwd$ and $onPolicy$ are coded 0/1.

5.2 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:

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

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, values were derived using Linear RL, estimated from a transition matrix $\mathbf{T}$ and rewards $\mathbf{r}$ (both learned over the course of the experiment):

$$\exp(\mathbf{v}_N^*/\lambda) = \mathbf{M}\mathbf{T}_{NT} \exp(\mathbf{r}_T/\lambda) \tag{5}$$

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

$$\mathbf{M} = (\mathrm{diag}(\exp(-\mathbf{r}_N/\lambda)) - \mathbf{T}_{NN})^{-1} \tag{6}$$

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 delta learning rule:

$$T[island, boat] = \alpha T 1_{boat=c} + (1 - \alpha T) T[island, boat] \tag{7}$$

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:

$$P(i_t = i \in L, R) \propto \exp(\beta_{island} V(i) + \beta_{persistenceI} LastChosen(i)) \tag{8}$$

5.3 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 (Huys et al., 2011). 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 (Oakes, 1999), itself approximated using the Hessian of a single Newton-Raphson step.

5.4 Simulation

To test whether agents with high and low values of λ exhibited interactions indicative of MB or temporally-abstracted behavior, we simulated 200 agents per condition. Simulated agents followed the Linear RL behavioral choice model outlined above Each agent was simulated on a randomly generated trial sequence following the same block reward structure as human data. Softmax temperatures were drawn from a normal distribution parameterized identically to the group-level fit on human data. To isolate the interaction terms from history effects, we set the learning rates (both for reward as well as transitions) equal to 1, such that the estimates of internal state being reflected in prior choice as well as prior observed reward were in fact equal to those values.

5.5 Model Comparison

To test whether human choice behavior was best-explained by a mixture of MB and SR planning, rather than either alone, we performed a nested model comparison, contrasting a model where λ, controlling the the balance mix, against a model where $\log(\lambda)$ was fixed at either -5 or +5, which respectively lead to MB or SR behavior. In order to compare models with different numbers of free parameters, correcting for any bias due to overfitting, we computed a cross-validated approximation to the negative log marginal likelihood for each subject. Specifically, we used leave-one-session-out cross validation for the population-level prior parameters and a Laplace approximation for the per-subject parameters: for each subject, we refit the population-level model omitting that session, then conditional on that prior, we computed a Laplace approximation to that subject's log marginal likelihood, obtaining a score for each subject and model. Finally, we use paired tests on these scores across subjects, between models, to formally test whether any model fit consistently better over the population.

5.6 Blockwise Change Analysis

To estimate whether subjects shifted their strategies after congruent versus incongruent block changes, we fit a two-step choice model, with island values calculated with Linear RL (see above)

but with two independent values of λ, termed $\lambda+$ and $\lambda-$. $\lambda+$ was used to estimate island values throughout a block following a congruent reward change, whereas $\lambda-$ was used to estimate island values throughout a block following an incongruent block change. Note that 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/`.

6 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.

References

Bates, D., Alday, P., Kleinschmidt, D., José Bayoán Santiago Calderón, P., Zhan, L., Noack, A., … Nazarathy, Y. (2023, September). *Juliastats/mixedmodels.jl: v4.22.1.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.8357196` doi: 10.5281/zenodo.8357196

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. Retrieved from `https://epubs.siam.org/doi/10.1137/141000671` doi: 10.1137/141000671

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009, December). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280. doi: 10.1016/j.cognition.2008.08.011

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011, March). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. doi: 10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005, December). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. doi: 10.1038/nn1560

Dayan, P. (1993, July). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, *5*(4), 613–624. doi: 10.1162/neco.1993.5.4.613

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023, May). jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software*, *8*(85), 5351. doi: 10.21105/joss.05351

Ekman, M., Kusch, S., & de Lange, F. P. (2023, February). Successor-like representation guides the prediction of future events in human visual cortex and hippocampus. *eLife*, *12*, e78904. doi: 10.7554/eLife.78904

Everitt, B. J., & Robbins, T. W. (2005, November). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. doi: 10.1038/nn1579

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017, April). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, *6*, e17086. doi: 10.7554/eLife.17086

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016, March). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, *5*, e11305. doi: 10.7554/eLife.11305

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011, April). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS computational biology*, *7*(4), e1002028. doi: 10.1371/journal.pcbi.1002028

Keramati, M., Dezfouli, A., & Piray, P. (2011, May). Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. *PLOS Computational Biology*, *7*(5). doi: 10.1371/journal.pcbi.1002055

Kool, W., Gershman, S. J., & Cushman, F. A. (2017, September). Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, *28*(9), 1321–1333. doi: 10.1177/0956797617708288

Lynn, C. W., Kahn, A. E., Nyema, N., & Bassett, D. S. (2020, May). Abstract representations of events arise from mental errors in learning and memory. *Nature Communications*, *11*(1), 2313. doi: 10.1038/s41467-020-15146-7

Mattar, M. G., & Daw, N. D. (2018, November). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609–1617. doi: 10.1038/s41593-018-0232-z

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017, September). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. doi: 10.1038/s41562-017-0180-8

Oakes, D. (1999). Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 479–482. doi: 10.1111/1467-9868.00188

Piray, P., & Daw, N. D. (2020, May). Linear reinforcement learning: Flexible reuse of computation in planning, grid fields, and cognitive control. *bioRxiv*, 856849. doi: 10.1101/856849

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017, September). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Computational Biology*, *13*(9), e1005768. doi: 10.1371/journal.pcbi.1005768

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2021, August). *Neural evidence for the successor representation in choice evaluation.* bioRxiv. doi: 10.1101/2021.08.29.458114

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8. doi: 10.1002/hipo.22523

Seow, T. X. F., Benoit, E., Dempsey, C., Jennings, M., Maxwell, A., O'Connell, R., & Gillan, C. M. (2021, July). Model-Based Planning Deficits in Compulsivity Are Linked to Faulty Neural Representations of Task Structure. *Journal of Neuroscience*, *41*(30), 6539–6550. doi: 10.1523/JNEUROSCI.0031-21.2021

Simon, D., & Daw, N. (2011). Environmental statistics and the trade-off between model-based and TD learning in humans. *Advances in Neural Information Processing Systems*, *24*.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017, November). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. doi: 10.1038/nn.4650

Sutton, R. S. (1988, August). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44. doi: 10.1007/BF00115009

Sutton, R. S. (1991, July). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, *2*(4), 160–163. doi: 10.1145/122344.122377

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. doi: 10.1037/h0061626

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