Trial-by-trial learning of successor representations in human behavior

Ari E. Kahn[1], Dani S. Bassett[2], and Nathaniel D. Daw[1,3]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[2]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Psychology, Princeton University, Princeton, NJ, USA

November 7, 2024

Abstract

Decisions in humans and other organisms depend, in part, on learning and using models that capture the statistical structure of the world, including the long-run expected outcomes of our actions. One prominent approach to forecasting such long-run outcomes is the successor representation (SR), which predicts future states aggregated over multiple timesteps. Although much behavioral and neural evidence suggests that people and animals use such a representation, it remains unknown how they acquire it. It has frequently been assumed to be learned by temporal difference bootstrapping (SR-TD(0)), but this assumption has largely not been empirically tested or compared to alternatives including eligibility traces (SR-TD($\lambda > 0$)). Here we address this gap by leveraging trial-by-trial reaction times in graph sequence learning tasks, which are favorable for studying learning dynamics because the long horizons in these studies differentiate the transient update dynamics of different learning rules. We examined the behavior of SR-TD(λ) on a probabilistic graph learning task alongside a number of alternatives, and found that behavior was best explained by a hybrid model which learned via SR-TD(λ) alongside an additional zeroth-order predictive model. The relatively large λ we estimate indicates a predominant role of eligibility trace mechanisms over the bootstrap-based chaining typically assumed. Our results provide insight into how humans learn predictive representations, and demonstrate that people simultaneously learn the SR alongside lower-order predictions.

Introduction

Decisions in humans and other organisms depend, in part, on learning and using models that capture the statistical structure of the world, including the long-run expected outcomes of our actions. A range of neural and behavioral results (Schapiro et al. 2016; Garvert, Dolan, and Behrens 2017; Stachenfeld, Botvinick, and Gershman 2017; Momennejad et al. 2017; Russek et al. 2021; Ekman, Kusch, and de Lange 2023; Kahn et al. 2024) suggest that the brain uses temporally abstract representations like the successor representation (SR) (Dayan 1993), which predict future events aggregated over multiple steps. However, such long-range predictions are challenging to learn, and it remains unknown how the brain accomplishes this learning.

Although the SR has often been studied via its effects on choices in simple sequential decision tasks (Momennejad et al. 2017; Kahn and Daw 2024), trial-by-trial learning dynamics are difficult to study in this setting due to short temporal horizons and the indirect observation of predictions by way of choices. One setting in which it may be particularly promising to investigate such learning rules is stimulus-by-stimulus reaction times (RTs) in graph learning tasks, which reflect predictions more directly and over long temporal horizons. In particular, previous work has shown that in such tasks, people exhibit systematic RT biases. Among these is sensitivity to modular structure, which manifests as a 'cross-cluster surprisal effect' where RTs are heightened when transitioning between densely interconnected clusters of nodes (Karuza et al. 2017; Kahn et al. 2018). RT facilitation when transitioning from one state to another has been interpreted as reflecting learned expectancy, but the cross-cluster effect implies that these expectancies are not captured by simple one-step conditional probabilities (which are matched on cluster transition). Instead, these behavioral biases were shown to be well captured by a maximum-entropy prediction model that balances representational complexity with accuracy (Lynn et al. 2020a), and which is mathematically equivalent to the SR. The SR's predictions are higher within than between clusters, because multi-step transitions move

between clusters relatively rarely. These results suggest that peoples' learned representations of graph structure (as revealed by predictive facilitation of their RTs) reflect an SR.

While these effects occur in block-averaged behavior, and are predicted asymptotically by a well-learned SR, there are a number of distinct hypothesized mechanisms for how the SR might be learned in a trial-by-trial fashion (Russek et al. 2017). Such learning is computationally challenging and the mechanisms supporting it are of some interest, because it requires tracking long-run connections between multiple states—much like other cases of credit assignment but over multiple destination states simultaneously.

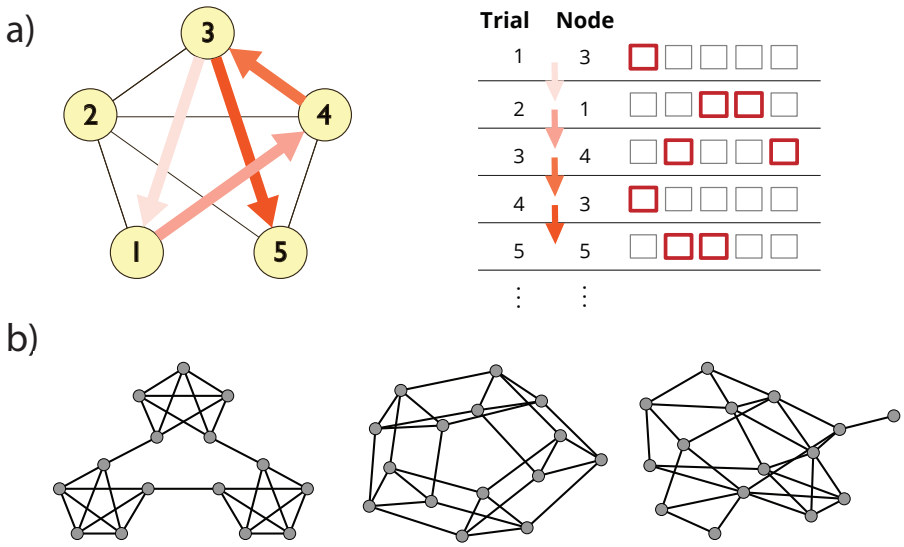

Figure 1: **SR Prediction on a Graph Learning Task. a)** Schematic of task design. Participants responded to a sequence of 1500 stimuli derived from a random walk on a graph, where each trial required pressing a one- or two-button combination of keys to identify the presented stimulus as quickly as possible. **b)** Construction of stimulus sequences. Walks were generated from either a modular (left), lattice (center), or random graph (right).

There are multiple candidate mechanisms for learning rules that support long-run forecasting. Since the first article introducing the SR (Dayan 1993), it has been widely assumed that the SR is directly learned via temporal-difference learning ("SR-TD"), specifically TD(0) bootstrapping (Dayan 1993; Gershman et al. 2012; Gardner, Schoenbaum, and Gershman 2018). TD(0) learns long-run predictions by serially chaining single-step "bootstrap" operations, backing up the prediction vector for each newly encountered state to its immediate predecessor and thereby spanning longer gaps via repeated traversals. Additionally, although less explored in this context, TD bootstrapping can be augmented or replaced by additionally maintaining "eligibility traces" over recently encountered states, allowing updates directly to span longer temporal gaps (SR-TD($\lambda > 0$)). A potential role for eligibility traces is intriguing in part because eligibility traces can also operate independently alongside or instead of bootstrapping (e.g., SR-TD(1) itself approximates Monte Carlo learning which can also be accomplished by combining eligibility traces with a simpler, biologically plausible Hebbian update), and because there have been recent suggestions that trace-driven learning might be instantiated in the brain using backward trajectory replay (Foster and Wilson 2006; Krausz et al. 2023; Sagiv et al. 2024). Further, the entropy-regularized prediction model developed

by Lynn et al. (2020a) suggests a learning rule that also, at least in expectation, corresponds to Hebbian learning augmented with eligibility traces. Importantly, although trace- and bootstrap-based learning rules converge to the same asymptotic representation of future occupancy, they do so via distinct trial-by-trial dynamics.

In the current study, we empirically test the fits of SR learning algorithms to the trial-by-trial dynamics of RTs in sequential graph learning tasks, by analyzing learning in three graphs from Kahn et al. (2018). If RTs on these tasks reflect long-run expectancy (Lynn et al. 2020b), then transient fluctuations in learned SRs due to variation in sequences should co-vary with biases in reaction times, allowing us to characterize the learning rules driving peoples' representations. Additionally, the update methods of these algorithms (in particular, bootstrapping and traces) should themselves have distinct signatures observable in reaction times, by comparing transitions that either do or do not affect either mechanism.

Results

Successor Representation

The SR is a world model that enables planning via cached long-run predictions of future state. That is, it learns to predict $M(s, s') = \sum_{t=0}^{\infty} \gamma^t p(s_t = s' | s_0 = s)$, a temporally discounted future occupancy of state s' given state s. In a tabular setting, this can be represented via an $n \times n$ matrix for n discrete states, where $M_{s,s'} = M(s, s')$. By coupling this model with a one-step estimated reward at each state, $R(s)$, the long-run value of a state can be calculated as the dot product $V(s) = M_{s,:}R$. (One-step estimated rewards can be trivially learned, for example, with a delta rule.)

Candidate learning rules for the SR itself, however, are more varied, notably featuring two distinct mechanisms for bridging temporal gaps: serially chaining one-step outcomes via bootstrapping and broadcasting them via eligibility traces. These two mechanisms, in turn, represent special cases of the SR-TD(λ) algorithm, which interpolates between them (see Methods). While these rules will, in appropriate conditions, converge asymptotically to the same matrix M, their transient behavior prior to (and indeed after) convergence is quite different, exhibiting unique biases due to bootstrapping and trace updates. By using RTs to read out internal predictions, we hoped to test which (or what combination) of these learning rules would best reflect recent trial history in subject behavior on a graph learning task.

Zeroth-Order Learning Model Best Predicts Reaction Times

What learning model best captures the transient trial-by-trial reaction time biases observed in graph learning tasks? Among those proposed (Russek et al. 2017), a plausible range of possibilities are parameterized by SR-TD(λ), which learns M via temporal difference learning. It does so by computing a bootstrapped TD error $\delta = I_j + \gamma M_{j,:} - M_{i,:}$ on each trial, and updating previous states based on this error via a trace whose decay is parameterized by λ. SR-TD(λ) thus encapsulates a family of models, where $\lambda = 0$ only updates the immediately preceding state via the bootstrap, and $\lambda = 1$ maintains an extended history and is comparable to a Monte Carlo update. Indeed, although most previous work has assumed $\lambda = 0$, eligibility traces help to speed learning over long temporal horizons. Further, arguably the most biologically plausible case of all is $\lambda = 1$, since this case approximates Hebbian learning with eligibility traces. In contrast, the bootstrap rule for $\lambda < 1$ requires a high-dimensional vector of error signals spanning destination states or state features, which imposes connectivity requirements that have, for some proposals, been argued to

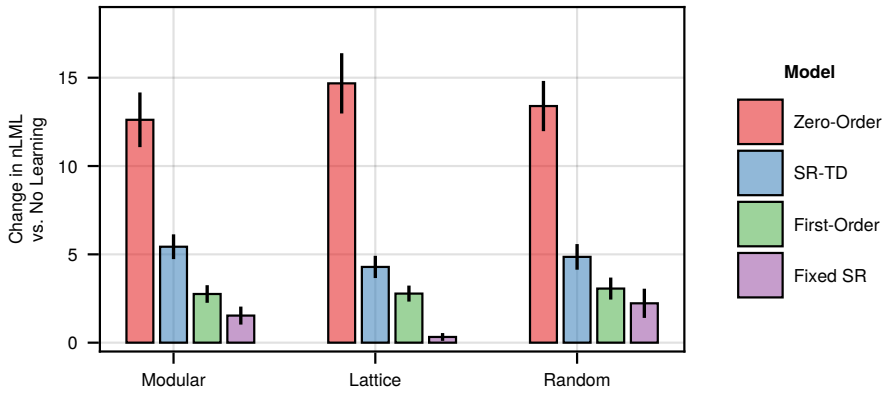

Figure 2: **Learning Models.** Model fits for each subject group. Indicated is the per-subject improvement in approximated negative log marginal likelihood versus a baseline model which performs no learning, where a higher value indicates a better fit. Error bars indicate +/- 1.96 standard error.

be implausible (Gardner, Schoenbaum, and Gershman 2018; Lau, Monteiro, and Paton 2017; Lee et al. 2024).

We first compared fits to trial-by-trial reaction times by SR-TD and a number of simpler alternatives. These included a fixed SR without trial-by-trial learning (based on expected transition probabilities across the entire walk), a model that learned first-order transition probabilities (i.e., the one-step state-to-state transition matrix T), and a zeroth-order learning rule that simply learned destination state probabilities unconditional on the predecessor state. Reaction times were fit via a shifted log-normal distribution, $\log(rt_t - shift) \sim N(\mu_t, \sigma^2)$ where rt_t is the reaction time on trial t, $shift$ is the minimum possible RT predicted by the model (fit as a fraction between 0 and 1 of that subject's minimum observed RT), and μ_t is a predicted value for trial t given by one of the respective models, which in the case of the SR was assumed to be proportional to its future occupancy weight M_{s_{t-1},s_t} plus the weighted sum of a number of additional nuisance variables (see Methods). In order to compare the dynamics of different learning rules while excluding the possibility that these might interact with arbitrary assumptions about initialization of M, the first 500 trials of each walk were used to equilibrate the model and to establish model representations, but did not contribute to likelihood. Note that due to stochastic state encounters, all the learning models predict different patterns of fluctuations around the learned representation even in the steady state, and it is these we wish to examine to isolate learning dynamics separate from initialization.

We found that the SR-TD model provided a significantly better fit than did the first-order model (p-values $< 1 \times 10^{-19}$) to all sets of subjects, suggesting that, in-line with previous analyses, reaction times are reflective of the SR. Moreover, the trial-by-trial learning model provided a significantly better fit than did the SR without learning: fits which incorporated trial-by-trial learning outperformed those with a static, ground-truth knowledge of the underlying graph structure (p-values $< 1 \times 10^{-8}$), indicating that we are capturing trial-by-trial transients even after the equilibration period. To our surprise, however, the zeroth-order model outperformed SR-TD (p-values $< 1 \times 10^{-21}$). In other words, a model learning unconditional recency-weighted node frequency explained more variance in trial-by-trial RTs than any of the higher-order models, raising a

question of whether people were in fact learning the SR. This was the case even though, following earlier work, the SR models incorporated nuisance variables which were included to control for such recency (but were evidently insufficiently flexible to do so; see Supplement).

SR Learning Persists Alongside Zeroth-Order Learning

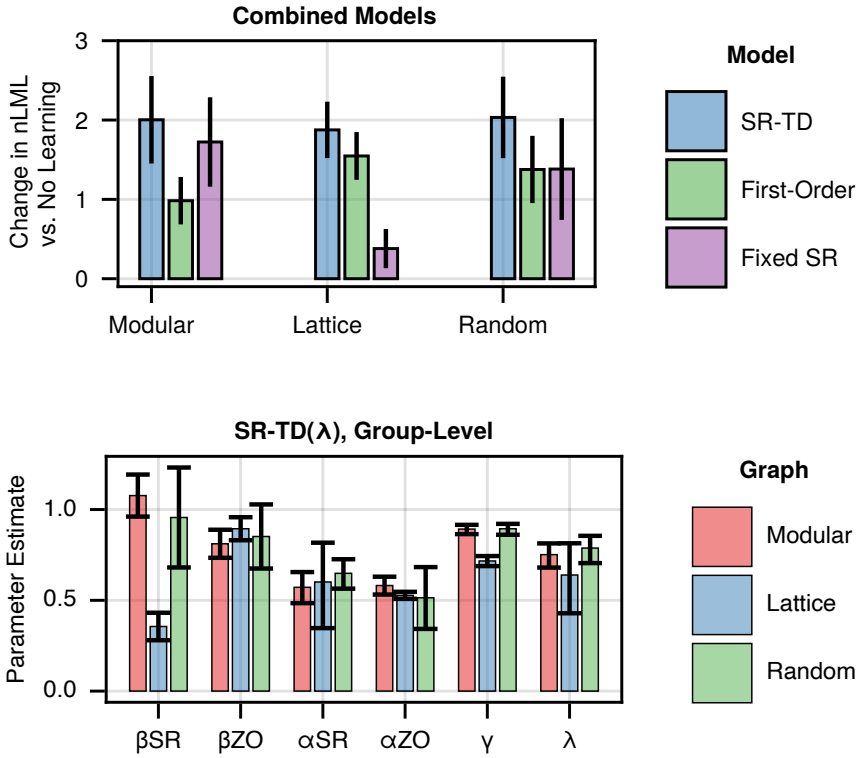

Figure 3: **Combined Learning Models. Top)** Model fits for combined models for each graph. Indicated is the per-subject improvement in approximated negative log marginal likelihood versus the zeroth-order model, where a higher value indicates a better fit. Error bars indicate +/- 1.96 standard error. **Bottom)** Group-level parameter estimates for the best fitting model, which combined zeroth-order and SR-TD(λ) learning. Error bars indicate 95% CI. SR: Softmax temperature and learning rate for the SR-TD(λ) component. ZO: Softmax temperature and learning rate for the zeroth-order component.

Next, we asked whether behavior was best explained by zeroth-order learning alone, or whether further higher-order learning effects were present alongside it. Converging evidence via other approaches suggest that people learn predictive representations in such contexts (Schapiro et al. 2016; Garvert, Dolan, and Behrens 2017; Tacikowski et al. 2024), and thus, a likely explanation is that recency-driven expectations (e.g., unconditional prior probabilities over states) exist alongside higher-order state-to-state conditional representations, with differing degrees of impact on reaction times. We thus combined the models tested previously, so as to simultaneously (but independently) learn zeroth-order predictions alongside higher-order effects, with different learning

rates and softmax βs for the two components. We then compared the fit of these models to one another, as well as to the model of zeroth-order learning alone.

We observed that models which included higher-order predictions displayed improved fits across all subject groups, when compared to the zeroth-order model. Fits of the SR-TD model consistently provided the best fit in all subject groups, with differences from the first-order model being statistically significant in all subject groups (p-values < 0.0002), and differences from the zeroth-order-alone model being statistically significant in all subject groups as well (p-values $< 1 \times 10^{-10}$). Additionally, the difference from an SR model without learning was significant for lattice and random graphs. Thus, while zeroth-order learning drives behavior more strongly than the higher order effects, it does not fully explain reaction time biases. Instead, those biases are best explained by a combination of the zeroth-order learning and a learned multi-step predictive model.

We next examined the fit parameters to understand the contribution of different learning mechanisms to our results. In addition to a softmax temperature β that scales the contribution of the SR expectancies onto reaction times, SR-TD is parameterized by γ, which controls how far into the future predictions extend, and λ which controls the duration of eligibility traces. Two special cases are of particular interest. First, $\lambda = 0$ (bootstrapping alone) is the case most often assumed previously. At the other extreme, $\lambda = 1$ relies heavily on long-lasting eligibility traces (and indeed admits of Monte Carlo variants that can be implemented with only Hebbian learning and no bootstrapping at all). We examined estimates from our combined model, and found that λ was consistently estimated quite high (modular: 0.752, 95% CI: [0.682, 0.812], lattice: 0.637, 95% CI: [0.448, 0.798], random: 0.786, 95% CI: [0.725, 0.838]) but that the extreme values of 0 and 1 were excluded. This suggests, in contrast to previous work, that learning is predominantly driven by eligibility traces, though we do not rule out bootstrapping altogether.

Human RTs Exhibit Signatures of Trace Updating but not Bootstrapping

We next sought to more directly detect model-agnostic signatures in the data of eligibility trace and bootstrap learning mechanisms. The key signature of trace learning is that when a state sequence such as ABC is visited, the prediction from A is immediately updated to expect C. (Conversely, bootstrapping without eligibility traces requires a second AB transition to chain expectation from C back to B then B back to A.) This formulation suggests that we could look for evidence of reaction times being modulated in this way, by finding sequences of trials that differ only by whether a trace update could have occurred, and testing whether reaction times systematically differ between those sequences.

In particular, we compared reaction times for the final elements of sequences of the form '$SXTST$' versus 'X_1X_2TST', which measures expectancy of a target T on its second visit, following the second visit to a source S (in the first case) or following the first visit to S (in the second case). If a trace is used for learning, then expectancy of T following S is increased by the initial sequence 'SXT', which is tested by the comparison of 'SXT' vs 'X_1X_2T' (Fig. 4A,C). We directly tested for this effect by finding all sequences matching the above patterns, and fitting a hierarchical mixed-effects model to subject reaction times (see Methods). As predicted, we observed that participants were significantly faster at responding to the final element of sequences of the first type ($\mu = 30$ms, SE=0.01, $z = 2.93$, $p < 0.0034$), suggesting that they updated their expectations in a trace-like manner (Fig. 4E, left).

Next, we tested a similar signature for a bootstrap backup, the central feature of TD(0). For this purpose, we contrasted reaction times from sequences matching '$BTSBST$' to those matching '$XTSBST$'. In the first sequence, the transition from a bootstrap node B to the target T at the start increases the prediction of T given B. The subsequent transition from S to B should, in turn,

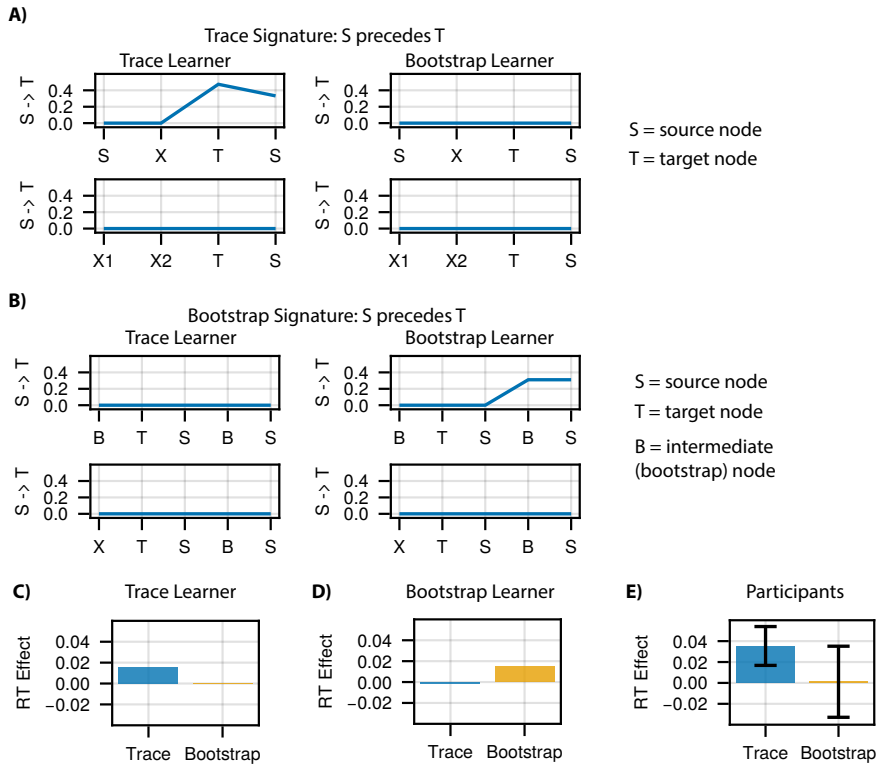

Figure 4: **Model Agnostic Signatures of Trace Updating and Bootstrapping. A)** The sequence $S \rightarrow X \rightarrow T \rightarrow S$ increases the weight given to $p(T|S)$ for trace learners but not bootstrap learners. The sequence $X_1 \rightarrow X_2 \rightarrow T \rightarrow S$ is provided as a recency-matched control. **B)** The sequence $B \rightarrow T \rightarrow S \rightarrow B \rightarrow S$ increases the weight given to $p(T|S)$ for bootstrap learners but not trace learners. The sequence $X \rightarrow T \rightarrow S \rightarrow B \rightarrow S$ is provided as a recency-matched control. **C)** Expected difference in RT between both sequences and their matched controls for a bootstrap learner. Bootstrap learners exhibit a significant effect of the bootstrap manipulation, but not of the trace manipulation. **D)** Expected difference in RT between both sequences and their matched controls for a trace learner. Trace learners exhibit a significant effect of the trace manipulation, but not of the bootstrap manipulation. **E)** Participant RTs exhibit a significant trace effect, but not a significant bootstrap effect. Error bars indicate +/- 1.96 standard error.

chain that prediction of T from B to S (an effect not occurring in the second sequence as the transition from B to T was never initially increased), and thus decrease reaction times for the final transition from S to T (Fig. 4B,D). However, we detect no significant effect of sequence type for reaction times here, in contrast to our trace effect (μ=3.5ms, SE=0.219, p<0.87, Fig. 4E, right). While this may be consistent with the high estimated λ from the model-based analysis, negative results should be interpreted with caution. Here, compared to the trace effect, the increased pattern length (six elements instead of five) means that we have fewer observations — 15,493 across all subjects, versus 94,000, limiting our power compared to the trace analysis. The standard error on the estimate (reflecting this reduced power) implies that this analysis cannot confidently rule out an effect on RTs roughly the size of that estimated for the trace.

Discussion

It has long been known that reaction times in sequential response tasks like graph learning are facilitated by learned expectations (Cleeremans and McClelland 1991; Bornstein and Daw 2012; Kakaei, Aleshin, and Braun 2021; Park et al. 2022; Solomon and Schapiro 2024). More surprisingly, it has recently been shown that these predictions span multiple future timesteps (Karuza et al. 2017; Kahn et al. 2018; Lynn et al. 2020a; Stiso et al. 2022; Ren, Zhang, and Luo 2022; Wittkuhn, Krippner, and Schuck 2022). In this respect, they coincide suggestively with a form of temporally abstracted prediction model called the successor representation, which has also been a focus of recent study in other areas of decision making (Momennejad et al. 2017; Russek et al. 2021; Ekman, Kusch, and de Lange 2023; Wientjes and Holroyd 2024; Kahn and Daw 2024), cognitive map learning (Stachenfeld, Botvinick, and Gershman 2014; Whittington et al. 2022; Garvert et al. 2023; Tacikowski et al. 2024; Yoo, Chrastil, and Bornstein 2024), and memory (Wittkuhn, Krippner, and Schuck 2022; Zhou et al. 2023). These RT effects thus might offer a novel window into the mechanisms by which such long-run predictions are formed over delays, which is one of the key computational challenges in making effective decisions but challenging to study in many of the other settings where the SR arises. Here we investigate the trial-by-trial learning rules driving predictive effects in graph learning.

Across several different graph structures and using both model fits and model-agnostic signatures, we find evidence for a predominant role of eligibility traces in such learning. While not the rule most often assumed, these results are in line with prior suggestions that traces are a biologically plausible mechanism to enable the learning of predictive associations over delays (Foster and Wilson 2006; Sagiv et al. 2024; Lau, Monteiro, and Paton 2017).

Reaction times are strongly driven by zeroth-order learning effects alongside SR

One surprising aspect of the current results was that, although SR learning is significantly demonstrable, it exists alongside more basic learning effects. Specifically, RTs were strongly driven by zeroth-order learning, i.e., a vector of state probabilities derived from a running average over recent experience but unconditional on the predecessor state. In fact, considered in isolation, these zeroth-order learning effects provide a better prediction of RTs than do higher-level models alone. This was true even though our SR models contained nuisance variables intended to control for recency effects, which were evidently insufficiently flexible to capture the timescale of recency. Thus, properly accounting for these effects is crucial, particularly if we hope to characterize higher-order learning.

Moreover, this result suggests a more comprehensive model of learning than previously proposed. Of course, participants have no reason to assume that the states they encounter are drawn from a graph (i.e., Markov according to first-order conditional probabilities), but instead likely entertain multiple predictive models. In this respect, the zeroth-order model is the simplest case of the idea that people and animals learn predictive models across multiple timescales (Sutton 1995; Momennejad and Howard 2018), in this case no timescale at all. This ability to learn models across multiple timescales may be supported, in part, by range of timescales of dopaminergic signaling observed in animal studies (Schultz 2007; Bromberg-Martin et al. 2010; Mohebi et al. 2024).

Mechanisms for learning long-run predictions

The SR predicts aggregated, discounted future long-run encounters of each state (here, graph node) conditional on the current state. Such predictions are of particular interest in reinforcement learning,

because they may be used to forecast the long-run consequences of candidate actions and guide choices toward later rewarding states. In the present task, where movements are instructed rather than chosen, such predictions may more implicitly drive motor preparation, thereby facilitating faster RTs. However, in this setting it is less clear why long-run—rather than next-step conditional—expectancies should be used. This may reflect the brain learning predictions optimized to other, more general settings (e.g., non-Markovian dynamics or sequential choice tasks) rather than graph learning *per se*, or it may reflect the learning of a one-step model that is, in effect, regularized by longer-run expectancy (Lynn et al. 2020a).

Regardless, similar to the more studied case of temporal credit assignment (i.e., predicting long-run future rewards; Sutton 1984; Walsh and Anderson 2011; Gersch et al. 2014; Her et al. 2016; Hamid, Frank, and Moore 2021; Krausz et al. 2023), learning such predictions requires spanning temporal gaps to learn associations between graph nodes encountered at a delay. Also as in the case of rewards, three distinct, though not mutually exclusive, classes of mechanisms have generally been proposed for learning an SR (Russek et al. 2017): chaining predictions successively by bootstrapping, spanning temporal gaps by maintaining sustained eligibility traces, and repeatedly iterating a one-step model to increase the timescale of prediction.

In the present study, we focus on the first two of these (see below for more on iteration), which can be hybridized by a single parameterized learning rule, SR-TD(λ). This form is particularly convenient for descriptive estimation because it casts the mechanism question as that of estimating the parameter λ, enabling us to identify mixtures of both but also detect nested special cases $\lambda = 0$ corresponding to pure bootstrapping and $\lambda = 1$ similar to pure eligibility traces.

Most work on the SR has tended to envision that its predictions are learned by bootstrapping, i.e. SR-TD(0) (Wittkuhn, Krippner, and Schuck 2022; Garvert et al. 2023). Here, each time a state is encountered, its own forward-looking predictions are "backed up" to its immediate predecessor state, with a net effect that thus accumulates and chains backward over multiple traversals. This is accomplished using temporal-difference prediction errors, and the popularity of the algorithm is probably in part due to evidence that the brain's dopamine system appears to signal such bootstrap errors for the case of chaining reward expectancy (Enomoto et al. 2011; Maes et al. 2020; Amo et al. 2022). However, whereas reward is a scalar and has been argued to be appropriate for a relatively homogeneous dopamine projection, it is less clear whether dopamine (or if not, what other system) can deliver the higher-dimensional set of vector-valued prediction errors required to bootstrap individual SR predictions for many successor states (Gardner, Schoenbaum, and Gershman 2018; Lau, Monteiro, and Paton 2017).

A second mechanism for learning associations at a delay is to maintain "eligibility traces" for each state that increment when it is encountered and gradually decay over time (Sutton 1988). These can operate in isolation, standing in for presynaptic activity in a Hebbian learning rule to associate previously encountered states with each newly encountered one. They can also work in concert with bootstrapping; in particular, for $\lambda > 0$, SR-TD(λ) uses such traces to transport each bootstrap-based update over multiple preceding timesteps. In the extreme case of SR-TD(1), learning is essentially driven only by traces and similar to the simpler Hebbian rule, and the bootstrap (though still algebraically present) contributes only transiently and incidentally. Traces have been argued to be neurally plausible (for the SR, arguably moreso than the bootstrap), in that eligibility might be maintained by some persistent chemical change locally at a synapse or spine (Wickens 1990; Wickens and Kötter 1995; Houk, Adams, and Barto 1995). Alternatively, neurons representing eligible states might be reactivated after a delay, via history stored in some separate memory system (Foster and Wilson 2006). The maximum entropy graph learner envisioned by Lynn et al. (2020a) also works by reactivating recently active predecessor states, and corresponds (at least in expectation) to a trace update.

Accordingly, in our model fitting analyses, we focused on estimating λ (alongside other parameters and nuisance effects) to detect evidence for bootstrapping and/or traces. Best-fit values tended to be intermediate, though closer to 1 than 0. Taken at face value (since both extreme models are rejected), these results support a contribution of both mechanisms. In conjunction with evidence for traces also in our separate model-agnostic analyses (comparing RTs between different state sequences chosen to expose the effects of either bootstrap or traces), we take our results as clear evidence of a strong role for traces.

We are more cautious about the evidence in favor of bootstrapping, due to high (albeit sub-unity) estimated λ and also the failure of our model-agnostic analyses to detect positive evidence for it. Interestingly, this negative result stands in contrast to our recent clearer detection of bootstrapping using somewhat similar analyses (albeit of neural data) in a reward learning task, where TD(0) is also a more neurally plausible algorithm (Krausz et al. 2023). Of course, this difference may reflect lower power in the model-agnostic analyses for bootstrap relative to traces (due to fewer, longer sequences required to expose the former). Another possibility is that a better fitting trace-only learning rule remains to be discovered: SR-TD(1) can be learned through a variety of methods that rely on strictly Hebbian trace rules instead of bootstrapping. These variations in turn all exhibit distinct transients, and further work is required to investigate whether any of these might plausibly provide a better explanation of behavior than SR-TD(λ).

Finally, we also considered, but do not report here, "SR-MB" learning rules (Russek et al. 2017), which estimate a simple one-step conditional state-state probability model T via Hebbian updates, i.e., with neither traces nor bootstrap. These can be iteratively accumulated after each update (equivalent to a sum of matrix powers (Fang et al. 2023), or a matrix inversion $M = (I - \gamma T)^{-1}$) to recover the long-run SR (Sutton and Pinette 1985). This represents the third general route to long-run SR-prediction, analogous to iterative model-based rollouts in the reward-learning case. We omit these models from the analyses reported here, because they did not capture any unique variance in RTs, suggesting no detectable contribution of SR-MB. Instead, in all cases, we estimated a nested special case ($\gamma \to 1$) that is, for this model, equivalent to the zeroth-order learner.

Temporal discounting and choice

Alongside the parameter λ, which governs how long eligibility persists for learning over temporal delays, we also estimate a second parameter, γ, which controls how far into the future state occupancy predictions accumulate. In choice tasks, γ has long been of interest because it controls intertemporal preferences, i.e., how reward amounts trade off against delays. In the current task (which involves no choices), the parameter instead governs the timescale of implicit predictions as reflected in RT facilitation. Interestingly on the average over participants, we estimate a remarkably consistent degree of temporal discounting across all three graph topologies, with all values falling between $\gamma = .717$ and $\gamma = 0.894$, equivalent to a temporal horizon of 3.5 to 9.4 timesteps. An interesting question for future work is whether the predictive model we measure in RTs is also used for choice, e.g. if subjects learn a graph and are subsequently asked to plan routes or choose between states for reward (Bornstein and Daw 2013; Wientjes and Holroyd 2024). In this latent learning case, γ estimated from RTs might reflect individual differences that also play out in intertemporal choice.

Methods

We fit our models to data from Kahn et al. (2018), where participants completed 1500 trials of a serial reaction time-like task, responding to a cue shown on the screen as rapidly as possible with a

one- or two-button key combination. Unbeknownst to participants, the sequence of cues followed a traversal through one of various graph structures: either a modular graph composed of 15 nodes in 3 clusters, a lattice graph, or a randomly connected graph. Each subject was assigned two of the three graphs, and performed the task twice in sequence with counterbalanced ordering. An additional subset of subjects performed the task on one of the three aforementioned graphs, followed by 500 trials on a fully connected graph. We include the first set of trials for those subjects, but exclude the fully connected graph from our analysis. We excluded all trials on which subjects responded incorrectly, as well as trials with implausible RTs (those less than 30 ms).

To compare proposed learning algorithms that would give rise to observed biases in reaction times, we estimated the likelihood of the sequence of RTs given the respective learning algorithm under a shifted log-normal distribution, which takes the form

$$\log(rt_t - shift) \sim N(\mu_t, \sigma^2), \tag{1}$$

where rt_t is the reaction time on trial t, $shift$ is the minimum possible RT predicted by the model (fit as a fraction between 0 and 1 of that subject's minimum observed RT), and μ_t is a predicted value for trial t given by one of the respective models. Per-subject parameters were fit hierarchically over the group via expectation maximization, and models were compared via approximated negative log marginal likelihood. To focus solely on transient trial-by-trial variation in prediction, and avoid artifacts reflecting the initialization of the models, likelihood was maximized only for the last 1000 trials in each sequence—the first 500 trials were simulated, but the likelihoods of the associated reaction times were not included for fitting purposes.

Zeroth-Order Model

In our Zeroth-Order model, reaction times were a function of individual node probabilities, where μ_t was defined as

$$\mu_t = \mu_0 + \beta_{trial} t + target_{2-15} + \beta_A W[s_t] + \beta_{ntrials} r_{ntrials} + \beta_{lag10} r_{lag10}, \tag{2}$$

where μ_0 is a baseline RT, *trial* is trial number to account for non-specific learning effects, $target_{2-15}$ are fits for the various finger combinations, and $W[s_t]$ was a function of the upcoming node. Each entry of W was initialized to 1/15, and W was updated on each trial as follows:

$$W \leftarrow (1-\alpha)W + \alpha I(s_t), \tag{3}$$

where $I(s_t)$ is a one-hot vector at position s_t.

Additionally, recency was controlled in-line with previous work by adding regressors $\beta_{ntrials}$ and $beta_{lag10}$, which respectively were applied to the number of trials elapsed since node s_t last occurred, and the number of occurrences of node s_t within the last 10 trials.

First-Order Model

In our first-order transition model, reaction times were a function of pairwise transition probabilities, where μ_t was defined as

$$\mu_t = \mu_0 + \beta_{trial} t + target_{2-15} + \beta_A T[s_{t-1}, s_t] + \beta_{ntrials} r_{ntrials} + \beta_{lag10} r_{lag10}, \tag{4}$$

where μ_0 is a baseline RT, *trial* is trial number to account for non-specific learning effects, $target_{2-15}$ are fits for the various finger combinations, and $T[s_{t-1}, s_t]$ was the entry of the learned transition

matrix from the previous to upcoming node. Each entry of T was initialized to 1/15, and T was updated on each trial as follows:

$$T[s_{t-1}, :] \leftarrow (1-\alpha)T[s_{t-1}, :], \tag{5}$$
$$T[s_{t-1}, s_t] \leftarrow T[s_{t-1}, s_t] + \alpha. \tag{6}$$

SR-TD Model

Our SR-TD model learned the SR via a TD(λ) update rule, and reaction times were a function of the predicted future occupancy of node s_{t+1} given node s_t. μ_t was defined as

$$\mu_t = \mu_0 + \beta_{trial} t + target_{2-15} + \beta_A M[s_{t-1}, s_t] + \beta_{ntrials} r_{ntrials} + \beta_{lag10} r_{lag10}, \tag{7}$$

where μ_0 is a baseline RT, *trial* is trial number to account for non-specific learning effects, $target_{2-15}$ are fits for the various finger combinations, and $M[s_{t-1}, s_t]$ was the SR's learned M matrix. M was initialized to $(I - \gamma T)^{-1} * T$ for a uniform transition matrix T, and updated on each trial as follows:

$$\begin{aligned}
&e[s_{t-1}] \leftarrow (1-\alpha)e[s_{t-1}] + 1 \\
&\delta = I[s_t] + \gamma M[s_t, :] - M[s_{t-1}, :] \\
&\text{for } s \text{ in 1:15} \\
&\quad M[s', :] \leftarrow \alpha e[s'] M[s', :] \\
&\text{done} \\
&T[s_{t-1}, s_t] \leftarrow T[s_{t-1}, s_t] + \alpha \\
&e \leftarrow \gamma \lambda e,
\end{aligned} \tag{8}$$

where discount factor γ and trace parameter λ were fit to each subject. Note that the SR as calculated here reflects a one-step-ahead variation, equivalent to $T + \gamma T^2 + \gamma^2 T^3 + \cdots$.

Static SR Model

Our static SR model calculated the SR for the empirical transition probabilities across the entire sequence for a given subject. That is, we calculated $T_{i,j} = \frac{\sum_t (s_{t-1}=i)(s_t=j)}{\sum_i (s_{t-1}=i)}$, and then $M = T * (I - \gamma T)^{-1}$. μ_t was defined as

$$\mu_t = \mu_0 + \beta_{trial} t + target_{2-15} + \beta_A M[s_{t-1}, s_t] + \beta_{ntrials} r_{ntrials} + \beta_{lag10} r_{lag10}, \tag{9}$$

where μ_0 is a baseline RT, *trial* is trial number to account for non-specific learning effects, $target_{2-15}$ are fits for the various finger combinations, and $M[s_{t-1}, s_t]$ was an entry of the static SR's M matrix.

Combined Models

For models that were a combination of Zeroth-Order and another learning model, μ_t was simply given as

$$\mu_t = \mu_0 + \beta_{trial} t + target_{2-15} + \beta_W W[s_{t-1}, s_t] + \beta_A A(s_{t-1}, s_t), \tag{10}$$

where $A(s_{t-1} + s_t)$ was M for the SR or T for the first-order model.

Model Comparison

To compare models which may have different numbers of free parameters, correcting for any bias due to overfitting, we marginalized the participant-level parameters by computing a Laplace approximation to the negative log marginal likelihood of each participant's data. Because the participant-level parameters were themselves estimated in the context of a hierarchical model governed by additional population-level parameters, we added an additional AIC penalty for these parameters, divided equally between subjects. This provides a fit score for each subject and model. Finally, we used paired t-tests on these scores across participants, between models, to formally test whether a model fit significantly better than another over the population (Stephan et al. 2009).

Trace Analysis

We sought to test in a model-agnostic fashion whether peoples' expectations reflected a trace update—that is, whether a transition expectation $S \to T$ was strengthened even when a third node X occurred between S and T as in $S \to X \to T$. To do so, we looked for sequences of the form 'SXTST' for any three distinct nodes S, X, and T, and took the reaction time on the final T as reflecting $S \to T$ after a trace update. We tested these reaction times against those from the final T of sequences of the form 'X_1X_2TST', in which recency of T is controlled for, but a trace update would not have occurred for S.

We then fit a hierarchical linear mixed model using *MixedModels* v4.24.1 (Bates et al. 2023) in *Julia* 1.10.4 (Bezanson et al. 2017) to these reaction times, where rt was predicted via

$$rt \sim \mu + \text{is_trace} + \text{target}_{2-15} + (1 + \text{is_trace} \mid \text{graphset}) + (1 + \text{is_trace} \mid \text{subject}), \tag{11}$$

where is_trace indicates RTs were from 'ABCAC' sequences (versus 'ABCDC'), $target_{2-15}$ are fits for the various finger combinations, and random effects were included for intercept and is_trace for each subject and graph instance (Modular/Lattice/Random, first or second walk). The model was fit across all subjects.

Bootstrap Analysis

Our model-agnostic test for a bootstrap update was performed similarly, but for a sequence designed to contrast bootstrap vs. non-bootstrap effects on reaction times. 'BTSBST' and 'XTSBST'. In 'BTSBST', the initial sequence 'BT' strengthens the prediction of T from B. For a bootstrapped update, 'SB' will then cause S's predictions to shift towards those of B, meaning 'ST' will be strengthened as well, leading to an enhanced RT on the final step of 'SR'. In the second sequence, the initial 'BT' is replaced by 'XT', and thus 'ST' should not be systematically upwardly biased by a bootstrap update.

We then fit a hierarchical linear mixed model using *MixedModels* v4.24.1 (Bates et al. 2023) in *Julia* 1.10.4 (Bezanson et al. 2017) to these reaction times, where rt was predicted via

$$rt \sim \mu + \text{is_bootstrap} + \text{target}_{2-15} + (1 + \text{is_bootstrap} \mid \text{graphset}) + (1 + \text{is_bootstrap} \mid \text{subject}), \tag{12}$$

where is_bootstrap indicates RTs were from 'ABCAC' sequences (versus 'ABCDC'), $target_{2-15}$ are fits for the various finger combinations, and random effects were included for intercept and is_bootstrap for each subject and graph instance (Modular/Lattice/Random, first or second walk). The model was fit across all subjects.

References

[1] Ryunosuke Amo et al. "A Gradual Temporal Shift of Dopamine Responses Mirrors the Progression of Temporal Difference Error in Machine Learning". In: *Nature Neuroscience* 25.8 (Aug. 2022), pp. 1082–1092. ISSN: 1546-1726. DOI: 10.1038/s41593-022-01109-2.

[2] Douglas Bates et al. *JuliaStats/MixedModels.jl: v4.22.1*. Version v4.22.1. Sept. 2023. DOI: 10.5281/zenodo.8357196.

[3] Jeff Bezanson et al. "Julia: A fresh approach to numerical computing". In: *SIAM Review* 59.1 (2017), pp. 65–98. DOI: 10.1137/141000671.

[4] Aaron M. Bornstein and Nathaniel D. Daw. "Cortical and Hippocampal Correlates of Deliberation during Model-Based Decisions for Rewards in Humans". In: *PLOS Computational Biology* 9.12 (Dec. 2013), e1003387. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003387.

[5] Aaron M. Bornstein and Nathaniel D. Daw. "Dissociating Hippocampal and Striatal Contributions to Sequential Prediction Learning". In: *European Journal of Neuroscience* 35.7 (2012), pp. 1011–1023. ISSN: 1460-9568. DOI: 10.1111/j.1460-9568.2011.07920.x.

[6] Ethan S. Bromberg-Martin et al. "Multiple Timescales of Memory in Lateral Habenula and Dopamine Neurons". In: *Neuron* 67.3 (Aug. 2010), pp. 499–510. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2010.06.031.

[7] Axel Cleeremans and James L. McClelland. "Learning the Structure of Event Sequences." In: *Journal of Experimental Psychology: General* 120.3 (1991), pp. 235–253. ISSN: 1939-2222. DOI: 10.1037/0096-3445.120.3.235.

[8] Peter Dayan. "Improving Generalization for Temporal Difference Learning: The Successor Representation". In: *Neural Computation* 5.4 (July 1993), pp. 613–624. ISSN: 0899-7667. DOI: 10.1162/neco.1993.5.4.613.

[9] Matthias Ekman, Sarah Kusch, and Floris P de Lange. "Successor-like Representation Guides the Prediction of Future Events in Human Visual Cortex and Hippocampus". In: *eLife* 12 (Feb. 2023). Ed. by Morgan Barense, Chris I Baker, and Helen Barron, e78904. ISSN: 2050-084X. DOI: 10.7554/eLife.78904.

[10] Kazuki Enomoto et al. "Dopamine Neurons Learn to Encode the Long-Term Value of Multiple Future Rewards". In: *Proceedings of the National Academy of Sciences* 108.37 (Sept. 2011), pp. 15462–15467. DOI: 10.1073/pnas.1014457108.

[11] Ching Fang et al. "Neural Learning Rules for Generating Flexible Predictions and Computing the Successor Representation". In: *eLife* 12 (Mar. 2023). Ed. by Srdjan Ostojic et al., e80680. ISSN: 2050-084X. DOI: 10.7554/eLife.80680.

[12] David J. Foster and Matthew A. Wilson. "Reverse Replay of Behavioural Sequences in Hippocampal Place Cells during the Awake State". In: *Nature* 440.7084 (Mar. 2006), pp. 680–683. ISSN: 1476-4687. DOI: 10.1038/nature04587.

[13] Matthew PH Gardner, Geoffrey Schoenbaum, and Samuel J Gershman. "Rethinking dopamine as generalized prediction error". In: *Proceedings of the Royal Society B* 285.1891 (2018), p. 20181645.

[14] Mona M Garvert, Raymond J Dolan, and Timothy EJ Behrens. "A Map of Abstract Relational Knowledge in the Human Hippocampal–Entorhinal Cortex". In: *eLife* 6 (Apr. 2017). Ed. by Lila Davachi, e17086. ISSN: 2050-084X. DOI: 10.7554/eLife.17086.

[15] Mona M. Garvert et al. "Hippocampal Spatio-Predictive Cognitive Maps Adaptively Guide Reward Generalization". In: *Nature Neuroscience* 26.4 (Apr. 2023), pp. 615–626. ISSN: 1546-1726. DOI: 10.1038/s41593-023-01283-x.

[16] Timothy M. Gersch et al. "Neural Correlates of Temporal Credit Assignment in the Parietal Lobe". In: *PLOS ONE* 9.2 (Feb. 2014), e88725. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0088725.

[17] Samuel J Gershman et al. "The successor representation and temporal context". In: *Neural Computation* 24.6 (2012), pp. 1553–1568.

[18] Arif A. Hamid, Michael J. Frank, and Christopher I. Moore. "Wave-like Dopamine Dynamics as a Mechanism for Spatiotemporal Credit Assignment". In: *Cell* 184.10 (May 2021), 2733–2749.e16. ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.03.046.

[19] Eun Sil Her et al. "Neuronal Activity in Dorsomedial and Dorsolateral Striatum under the Requirement for Temporal Credit Assignment". In: *Scientific Reports* 6.1 (June 2016), p. 27056. ISSN: 2045-2322. DOI: 10.1038/srep27056.

[20] James C. Houk, James L. Adams, and Andrew G. Barto. "A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement". In: *Models of Information Processing in the Basal Ganglia*. Computational Neuroscience. Cambridge, MA, US: The MIT Press, 1995, pp. 249–270. ISBN: 978-0-262-08234-1.

[21] Ari E. Kahn and Nathaniel D. Daw. "Humans Rationally Balance Mental Simulation and Temporally Abstract World Models". In: *Nature Communications Psychology* Accepted (Sept. 2024). DOI: 10.1101/2023.11.28.569070.

[22] Ari E. Kahn et al. "Network Constraints on Learnability of Probabilistic Motor Sequences". In: *Nature Human Behaviour* 2.12 (Dec. 2018), pp. 936–947. ISSN: 2397-3374. DOI: 10.1038/s41562-018-0463-8.

[23] Ari E. Kahn et al. "Network Structure Influences the Strength of Learned Neural Representations". In: *Nature Communications* Accepted (2024). DOI: 10.1101/2023.01.23.525254.

[24] Ehsan Kakaei, Stepan Aleshin, and Jochen Braun. "Visual Object Recognition Is Facilitated by Temporal Community Structure". In: *Learning & Memory* 28.5 (May 2021), pp. 148–152. ISSN: 1072-0502, 1549-5485. DOI: 10.1101/lm.053306.120.

[25] Elisabeth A. Karuza et al. "Process Reveals Structure: How a Network Is Traversed Mediates Expectations about Its Architecture". In: *Scientific Reports* 7.1 (Oct. 2017), p. 12733. ISSN: 2045-2322. DOI: 10.1038/s41598-017-12876-5.

[26] Timothy A. Krausz et al. "Dual Credit Assignment Processes Underlie Dopamine Signals in a Complex Spatial Environment". In: *Neuron* 111.21 (Nov. 2023), 3465–3478.e7. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2023.07.017.

[27] Brian Lau, Tiago Monteiro, and Joseph J Paton. "The Many Worlds Hypothesis of Dopamine Prediction Error: Implications of a Parallel Circuit Architecture in the Basal Ganglia". In: *Current Opinion in Neurobiology*. Computational Neuroscience 46 (Oct. 2017), pp. 241–247. ISSN: 0959-4388. DOI: 10.1016/j.conb.2017.08.015.

[28] Rachel S. Lee et al. "A Feature-Specific Prediction Error Model Explains Dopaminergic Heterogeneity". In: *Nature Neuroscience* 27.8 (Aug. 2024), pp. 1574–1586. ISSN: 1546-1726. DOI: 10.1038/s41593-024-01689-1.

[29] Christopher W. Lynn et al. "Abstract Representations of Events Arise from Mental Errors in Learning and Memory". In: *Nature Communications* 11.1 (May 2020), p. 2313. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15146-7.

[30] Christopher W. Lynn et al. "Human Information Processing in Complex Networks". In: *Nature Physics* (June 2020), pp. 1–9. ISSN: 1745-2481. DOI: 10.1038/s41567-020-0924-7.

[31] Etienne J. P. Maes et al. "Causal Evidence Supporting the Proposal That Dopamine Transients Function as Temporal Difference Prediction Errors". In: *Nature Neuroscience* 23.2 (Feb. 2020), pp. 176–178. ISSN: 1546-1726. DOI: 10.1038/s41593-019-0574-1.

[32] Ali Mohebi et al. "Dopamine Transients Follow a Striatal Gradient of Reward Time Horizons". In: *Nature Neuroscience* 27.4 (Apr. 2024), pp. 737–746. ISSN: 1546-1726. DOI: 10.1038/s41593-023-01566-3.

[33] I. Momennejad et al. "The Successor Representation in Human Reinforcement Learning". In: *Nature Human Behaviour* 1.9 (Sept. 2017), pp. 680–692. ISSN: 2397-3374. DOI: 10.1038/s41562-017-0180-8.

[34] Ida Momennejad and Marc W. Howard. "Predicting the Future with Multi-scale Successor Representations". In: *bioRxiv* (Oct. 2018), p. 449470. DOI: 10.1101/449470.

[35] Jungtak Park et al. "Reduced Functional Connectivity Supports Statistical Learning of Temporally Distributed Regularities". In: *NeuroImage* 260 (Oct. 2022), p. 119459. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2022.119459.

[36] Xiangjuan Ren, Hang Zhang, and Huan Luo. "Dynamic Emergence of Relational Structure Network in Human Brains". In: *Progress in Neurobiology* 219 (Dec. 2022), p. 102373. ISSN: 0301-0082. DOI: 10.1016/j.pneurobio.2022.102373.

[37] Evan M. Russek et al. *Neural Evidence for the Successor Representation in Choice Evaluation*. Aug. 2021. DOI: 10.1101/2021.08.29.458114.

[38] Evan M. Russek et al. "Predictive Representations Can Link Model-Based Reinforcement Learning to Model-Free Mechanisms". In: *PLOS Computational Biology* 13.9 (Sept. 2017), e1005768. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005768.

[39] Yotam Sagiv et al. *Prioritizing Replay When Future Goals Are Unknown*. Mar. 2024. DOI: 10.1101/2024.02.29.582822.

[40] Anna C. Schapiro et al. "Statistical Learning of Temporal Community Structure in the Hippocampus". In: *Hippocampus* 26.1 (2016), pp. 3–8. ISSN: 1098-1063. DOI: 10.1002/hipo.22523.

[41] Wolfram Schultz. "Multiple Dopamine Functions at Different Time Courses". In: *Annual Review of Neuroscience* 30.Volume 30, 2007 (July 2007), pp. 259–288. ISSN: 0147-006X, 1545-4126. DOI: 10.1146/annurev.neuro.28.061604.135722.

[42] Sarah H. Solomon and Anna C. Schapiro. "Structure Shapes the Representation of a Novel Category". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 50.3 (2024), pp. 458–483. ISSN: 1939-1285. DOI: 10.1037/xlm0001257.

[43] Kimberly L Stachenfeld, Matthew Botvinick, and Samuel J Gershman. "Design Principles of the Hippocampal Cognitive Map". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014.

[44] Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. "The Hippocampus as a Predictive Map". In: *Nature Neuroscience* 20.11 (Nov. 2017), pp. 1643–1653. ISSN: 1546-1726. DOI: 10.1038/nn.4650.

[45] Klaas Enno Stephan et al. "Bayesian Model Selection for Group Studies". In: *NeuroImage* 46.4 (July 2009), pp. 1004–1017. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2009.03.025.

[46] Jennifer Stiso et al. "Neurophysiological Evidence for Cognitive Map Formation during Sequence Learning". In: *eNeuro* 9.2 (Mar. 2022), ENEURO.0361–21.2022. ISSN: 2373-2822. DOI: 10.1523/ENEURO.0361-21.2022.

[47] Richard S. Sutton. "Learning to Predict by the Methods of Temporal Differences". In: *Machine Learning* 3.1 (Aug. 1988), pp. 9–44. ISSN: 1573-0565. DOI: 10.1007/BF00115009.

[48] Richard S. Sutton. "TD Models: Modeling the World at a Mixture of Time Scales". In: *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*. ICML'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., July 1995, pp. 531–539. ISBN: 978-1-55860-377-6.

[49] Richard S. Sutton and Brian Pinette. "The Learning of World Models By Connectionist Networks". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 7.0 (1985).

[50] Richard Stuart Sutton. "Temporal Credit Assignment in Reinforcement Learning". PhD thesis. University of Massachusetts Amherst, 1984.

[51] Pawel Tacikowski et al. "Human Hippocampal and Entorhinal Neurons Encode the Temporal Structure of Experience". In: *Nature* (Sept. 2024), pp. 1–8. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07973-1.

[52] Matthew M. Walsh and John R. Anderson. "Learning from Delayed Feedback: Neural Responses in Temporal Credit Assignment". In: *Cognitive, Affective, & Behavioral Neuroscience* 11.2 (June 2011), pp. 131–143. ISSN: 1531-135X. DOI: 10.3758/s13415-011-0027-0.

[53] James C. R. Whittington et al. "How to Build a Cognitive Map". In: *Nature Neuroscience* 25.10 (Oct. 2022), pp. 1257–1272. ISSN: 1546-1726. DOI: 10.1038/s41593-022-01153-y.

[54] J. Wickens. "Striatal Dopamine in Motor Activation and Reward-Mediated Learning: Steps towards a Unifying Model". In: *Journal of Neural Transmission / General Section JNT* 80.1 (Feb. 1990), pp. 9–31. ISSN: 1435-1463. DOI: 10.1007/BF01245020.

[55] Jeff Wickens and Rolf Kötter. "Cellular Models of Reinforcement". In: *Models of Information Processing in the Basal Ganglia*. Computational Neuroscience. Cambridge, MA, US: The MIT Press, 1995, pp. 187–214. ISBN: 978-0-262-08234-1.

[56] Sven Wientjes and Clay B. Holroyd. "The Successor Representation Subserves Hierarchical Abstraction for Goal-Directed Behavior". In: *PLOS Computational Biology* 20.2 (Feb. 2024), e1011312. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011312.

[57] Lennart Wittkuhn, Lena M. Krippner, and Nicolas W. Schuck. *Statistical Learning of Successor Representations Is Related to On-Task Replay*. Feb. 2022. DOI: 10.1101/2022.02.02.478787.

[58] Jungsun Yoo, Elizabeth R. Chrastil, and Aaron M. Bornstein. "Cognitive Graphs: Representational Substrates for Planning". In: *Decision* (2024), No Pagination Specified–No Pagination Specified. ISSN: 2325-9973. DOI: 10.1037/dec0000249.

[59] Corey Y. Zhou et al. *Episodic Retrieval for Model-Based Evaluation in Sequential Decision Tasks*. Feb. 2023. DOI: 10.31234/osf.io/3sqjh.