An Episodic Memory Model Can Account for Reinforcement Learning in Humans

Javier Masís[*,1], **Younes Strittmatter**[*,2] **& Jonathan D. Cohen**[1,2]
[1]Princeton Neuroscience Institute, Princeton University
[2]Department of Psychology, Princeton University
[*]equal contribution

Abstract

Multiple models have been proposed for how people solve reinforcement learning tasks, with the general conclusion that people may implement a mixture of "model-free" and "model-based" strategies. Here, we sought to investigate whether a process-based episodic memory model (called EGO) can account for human behavior in two inference tasks, the classic and revaluation 2-step tasks. The EGO model was able to mimic the behavior of humans and multiple RL models solely through its parametrization. While the parameters to mimic the same RL models across tasks differed, those to mimic human behavior were similar. These results suggest that rather than through a mixture of distinct processes, people may instead implement a single memory-based process to solve inference problems. The default parametrization of this process may reflect a global optimum for the average statistics of daily life, and may be adjusted through experience and control.

Keywords: reinforcement learning; episodic memory; working memory; context; 2-step task

Introduction

Reinforcement learning (RL) is concerned with a crucial function of cognition: predicting future states (Schultz et al., 1997; Sutton and Barto, 1998). RL models have been dichotomized into "model-free" (MF) and "model-based" (MB). People's behavior in a classic 2-step task resembles a mixture of MF and MB systems (Daw et al., 2011; Gillan et al., 2016), but in the similar revaluation 2-step task, the successor representation, a more sophisticated MF model, is required to explain people's behavior (Momennejad et al., 2017). This difference notwithstanding, it is unclear how these normative models would be implemented—and mixed—biologically. Latent cause inference (LCI) resolves the dichotomy, showing that MB systems can emerge from statistical MF systems when sufficient evidence from the environment warrants the creation of a new "cause" in the model's internal model of the world (Gershman and Niv, 2012; Gershman et al., 2015). As a Bayesian optimal model, however, it remains unclear how LCI could be implemented by a biological cognitive system.

Here, we implement the process-based, neurally plausible episodic memory generalization and optimization (EGO) framework (Giallanza et al., 2024), and apply it directly to two distinct RL tasks, the classic and revaluation 2-step tasks. We expand upon the qualitative fit to human behavior in the revaluation task in Giallanza et al. (2024) and show that EGO can account for RL behavior in humans across both tasks,

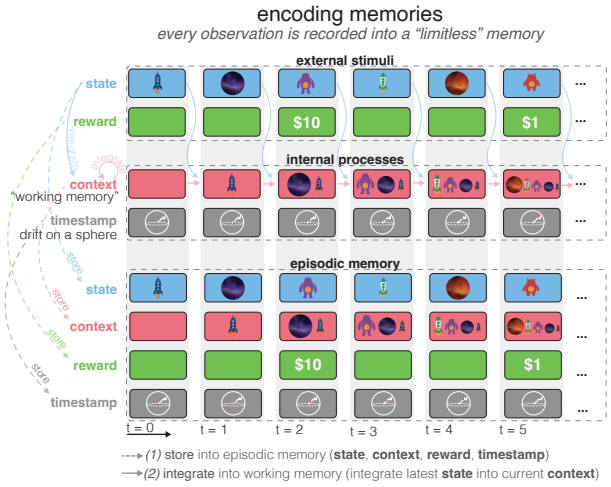

Figure 1: **Encoding memories.** On each time step, EGO: (1) stores a snapshot of external stimuli (state, reward) and internal processes (context, timestamp) into episodic memory and (2) integrates the latest state into the self-integrating context.

and can mimic MF and MB behavior solely through its parameterization, without the need for reinforcement, separate models, or a Bayesian optimal framework. Notably, behavior matching humans can be achieved in both tasks with a similar parameterization, but parameters that mimic MB and MF behavior across both tasks differ. This difference highlights the importance of process-based models over normative models, which by nature are tailored to provide the best solution to a very specific problem, and by extension are likely not to generalize. We argue that the model's modest elements: an episodic memory, a working memory, and a controller (amounting to a production system (Anderson, 1993; Lebiere & Anderson, 1993)) are reasonable and neurally plausible, and together provide a blueprint for a cognitive model that solves the general problem of prediction via inference.

Model

Encoding Memories

The EGO model (Giallanza et al., 2024) incorporates a limitless episodic memory that records "snapshots" of both external stimuli and internal processes at discrete time steps (Fig. 1), as well as a recurrent working memory of what the model has seen recently. External stimuli comprise states (i.e., sensory

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

stimuli), as well as any rewards received. Internal processes comprise a "working memory" (which we call context), as well as an internal clock that assigns a timestamp to every snapshot that is encoded into memory. At every time step during memory encoding, the model carries out two actions. First, it stores a copy of external stimuli and internal processes into its episodic memory. Second, it updates its working memory by integrating the latest state into the current context.

States & Rewards For ease of modeling, states are encoded as (orthogonal) one-hot vectors. For 8 possible states in a task, the first state is $\text{state}^1 = [0, 1, 0, 0, 0, 0, 0, 0]$, where, for convenience, the 0-indexed one-hot entry matches the state number. Rewards are encoded as scalars (e.g., 0 or 10).

Timestamps Timestamps are modeled as drift on an m-dimensional sphere. Timestamps that are near each other temporally, will have similar positions on the sphere. For example, if $\text{timestamp}_{t=n} = [-.14, .13, ..., -.01]$, then $\text{timestamp}_{t=n+1} = [-.15, .12, ..., -.02]$, whereas much later $\text{timestamp}_{t=n+100} = [3.02, 0.5, ..., -2.67]$.

Context Context is encoded as a vector matching the size of the state vector, preserving the correspondence between index and state (e.g., index 1 in the context vector also means the first state). Context updates according to a linear integrator at every time step t:

$$\text{context}_{t=n} = \gamma \text{state}_{t=n} + (1 - \gamma)\text{context}_{t=n-1} \qquad (1)$$

where the state integration rate γ determines the relative weight of the latest state versus the current context. If $\gamma = 1$, context becomes a copy of state one time step behind. As $\gamma \to 0$, context captures longer sequences of states. For example, let us say the agent has a $\gamma = 0.5$, starts with an empty $\text{context}_{t=n} = [0, 0, 0, 0, ...]$, and sees the sequence of states: $\text{state}^1_{t=n+1} = [0, 1, 0, 0, ...]$, $\text{state}^2_{t=n+2} = [0, 0, 1, 0, ...]$. Context would evolve as follows: $\text{context}_{t=n+1} = [0, 0.5, 0, 0, ...]$, $\text{context}_{t=n+2} = [0, 0.25, 0.5, 0, ...]$. Note that in encoding mode, this update to context would only happen *after* storing to episodic memory, such that the context stored at $t = n + 2$ is actually the context that was updated $t = n + 1$ (Fig. 1).

Retrieving Memories

Memories are retrieved via a two-step process: (1) query episodic memory with one or more desired fields (e.g., state, context, and timestamp), and (2) aggregate the query results to generate a retrieved memory (Fig. 2). When using context to query, it is generally first updated with the state of interest. This process allows context to "infer" future states and is further explained in the rollout section below.

Query Queries are similarity-based comparisons (e.g. dot product, cosine similarity) between a desired field (e.g., current state) and all the entries in that field in episodic memory (e.g., all recorded states). Let us say we would like to query memory with $\text{state}^1 = [0, 1, 0, 0, ...]$, and the

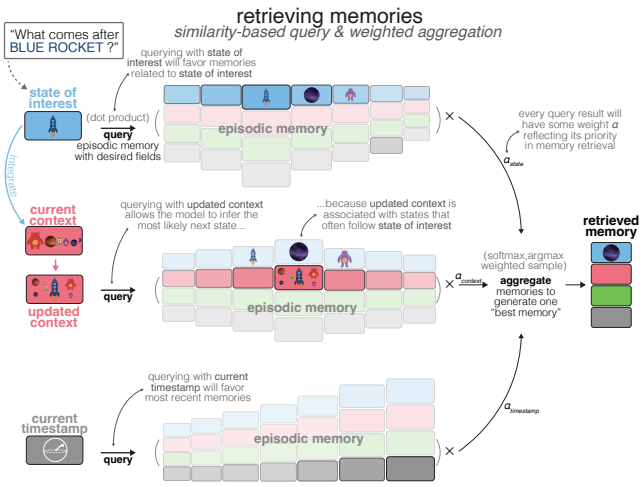

Figure 2: **Retrieving memories.** (1) Episodic memory is queried via a similarity-based comparison (e.g., dot product, cosine similarity) of one or more desired fields (e.g., state, context, timestamp) against all memories in those fields. When querying with context, a state of interest is first integrated into the current context, and the query is carried out with this updated context. (2) The similarity-ranked memories are aggregated (e.g., softmax, argmax, weighted average, weighted sampling) across fields to retrieve a "best" memory.

first four states in memory are $\text{state}^1_{t=1} = [0, 1, 0, 0, ...]$, $\text{state}^2_{t=2} = [0, 0, 1, 0, ...]$, $\text{state}^3_{t=3} = [0, 0, 0, 1, ...]$, $\text{state}^1_{t=4} = [0, 1, 0, 0, ...]$. We could write this as a matrix multiplication between the matrix of state memories $\mathbf{S}$, and the state of interest $\mathbf{s}_{query}$,

$$\text{query} = \mathbf{S}\mathbf{s}_{query} = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$$

where the result would be a vector containing the dot products (read: similarities) of every row with the state of interest. In this case, we would expect the first and fourth recorded states to be high matches.

Aggregate The query result will provide a similarity rank of every stored memory. The retrieved memory can then be chosen via multiple methods, such as softmax, argmax, weighted average, or a weighted sampling. Importantly, even though the similarity match is done in a field specific manner (e.g. state was queried above), the memories in other fields associated with the queried field (e.g. context, reward) will be assigned the same similarity result. When more than one field is queried (e.g., state, context, and timestamp), a weight is applied to the similarity results from each query (e.g. $\alpha_{state}, \alpha_{context}, \alpha_{timestamp}$), which is then used during the aggregation process. A greater weight will prioritize retrieval based on similarity to that particular memory field.

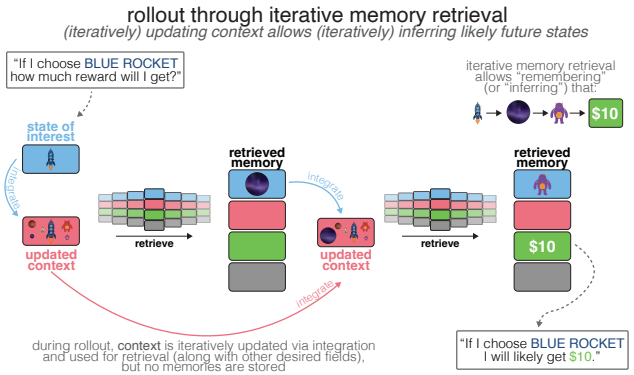

Figure 3: **Rollout.** First, a state of interest (e.g., blue rocket) is integrated into the current context. This updated context is used to query episodic memory and retrieve a memory. The retrieved state is integrated into the previously updated context, updating it again. This process is repeated until a desired endpoint is reached.

Rollout through Iterative Retrieval

Rollout is the iterative query and retrieval of memories to infer future states based on a starting state of interest (Fig. 3). Because context is a linear integrator of state, it can be used to query memory to predict future states.

Let us say that state^1 is always followed by state^2 is always followed by state^3, which is reliably paired with a reward of \$10, and that this sequence is reliably encoded in memory. Let us also say $\gamma = 1$ (Eqn. 1), making context a carbon copy of state one time step behind. We wish to predict states that will follow state^1, and whether they will lead to reward.

We first integrate our state of interest, state^1, into our current context. Because $\gamma = 1$, our updated context = state^1. In this simplified example, context is a carbon copy of state one time step behind, meaning that on time step $t = n + 1$, the recorded context will reflect the state at $t = n$, whereas the recorded state will be the current state at $t = n+1$. This means that when we query memory with context, the states associated with the most similar context memories will be one time step *ahead* of the state reflected in that context. With this in mind, we then query memory with our updated context = state^1 and retrieve a memory that has a state that has reliably occurred one time step ahead: state^2. We then integrate this retrieved state into our context, and our updated context = state^2. We query memory again and retrieve a state that has reliably occurred one time step ahead: state^3. We also retrieve the reward field associated with state^3: \$10. Thus, through rollout, we have effectively predicted the following sequence of states: state^1 → state^2 → state^3, and the associated reward of \$10.

Results

EGO model captures human & RL model behavior in classic 2-step task

Classic 2-step task The classic 2-step task (see Fig. 4 for task description) was designed to disambiguate MF and MB

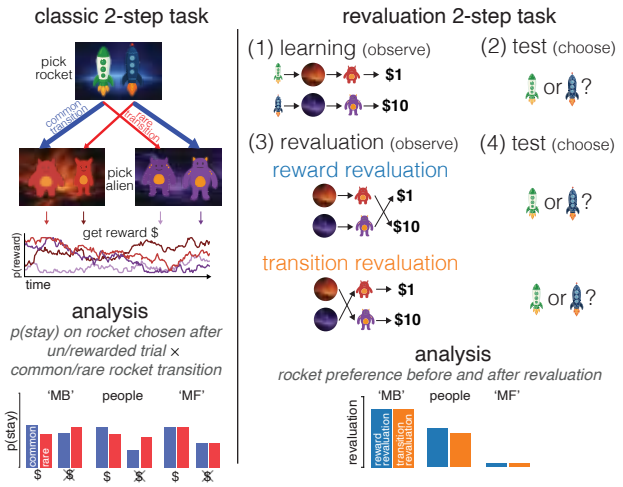

Figure 4: **Task descriptions. Classic 2-step task.** Participants pick one of two rocket ships. Each rocket ship leads to two planets with a high or low probability (common/rare transition). Participants then pick one of two aliens native to each planet. The aliens will reward participants according to independent drifting probabilities. **Revaluation 2-step task.** During a learning phase, participants will observe two fixed, randomly interleaved sequences of states (e.g., blue rocket → purple planet → purple alien (who always pays \$10)). They will then be asked which starting state (blue or green rocket) they prefer. During a revaluation phase, participants will observe modified sequences of states where either the rewards associated with each alien will flip (reward revaluation), or the planet where each alien resides will flip (transition revaluation). Notably, these sequences are truncated, starting at the second state, not the starting state. They will then be asked which starting state they prefer (blue or green rocket).

behavior by adding an intermediate step (which planet the chosen rocket lands on) between an agent's initial choice and the reward obtained. Whereas a MF agent will update its initial rocket choice's value based solely on whether it was ultimately rewarded, a MB agent will be sensitive to how it got to that planet (whether from a common or rare transition) and use this information to update the value of the more appropriate rocket. This behavior is measured by assessing the probability of repeating the previous rocket choice ("p(stay)") based on the previous trial's transition from rocket to planet (whether it was a common/rare transition), and whether it was ultimately rewarded. A MB agent will only show a main effect of the transition type, whereas a MF agent will only show a main effect of the reward outcome (Fig. 5a).[1] People show both effects, which has led researchers to describe human behavior as a mixture of MF and MB behavior (Daw et al., 2011).

[1]The choice function contains a free parameter to account for perseveration (tendency to repeat previous choice, first noted in monkeys in Lau and Glimcher, 2005), making p(stay)>0.5 across the board.

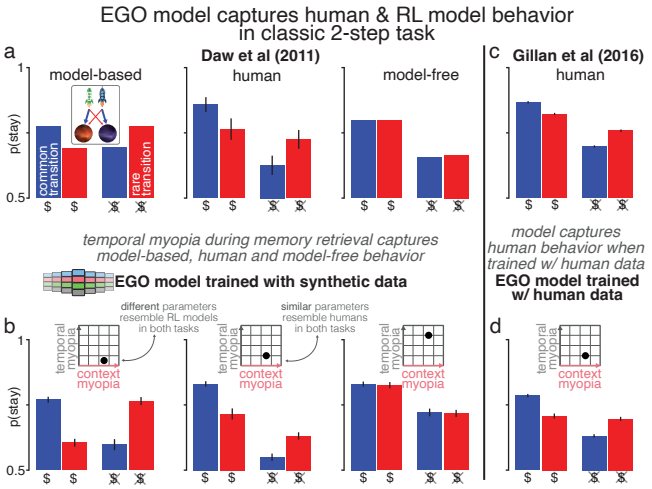

Figure 5: **EGO model captures human & RL model behavior in classic 2-step task.** **(a)** Cartoon of key results from Daw et al. (2011): human decision to repeat previous trial's choice (p(stay)) depends on previous trial's transition type and reward outcome, resembling a mixture of MF (only reward outcome matters) and MB behavior (only transition type matters). **(b)** EGO model trained with synthetic data: recency-bias during memory retrieval can capture MB (no bias), human (medium), and MF (large) behavior. **(c)** Human behavior from Gillan et al. (2016). **(d)** EGO model trained with human data in **c** captures human behavior with the same parameters as when trained with synthetic data.

Model training and choice process We first trained the EGO model on the classic 2-step task using synthetic data, and assessed its preference for repeating its previous rocket choice based on the previous trial's transition type and reward outcome. The synthetic data was generated by replicating the statistics of the original task (stable common/rare transition probabilities, independently drifting reward probabilities, 200 trials, 20 "participants"), but administering random choices (randomly choosing rockets and aliens). We first trained the model this way because it provided an unbiased sample of the task statistics, whereas training with human data would likely contain unaccounted-for biases, such as the known perseveration bias. We nevertheless trained the model with human data to ensure its results were consistent (see below, Fig. 5c, d).

Training in the EGO model is a qualitatively different process to training in a traditional RL model. In a traditional RL model, on each trial, the model explicitly updates its predicted values associated with each state based on an error signal. In the EGO model, on each trial, the model simply encodes in its episodic memory the states and rewards it has seen, but it does not have any explicit mechanism that associates or updates values to particular states outside of its episodic memory.

The choice process also differs across these model types. In a traditional RL model, the model "chooses" the state (e.g. rocket) with the higher associated predicted value.[2] In the EGO model, values associated with candidate choices are *implicit* in its episodic memory, and are accessed by the model querying its episodic memory with the candidate choices (states). When there is more than one step between the choice and the reward, it uses rollout (Fig. 3) to infer a candidate choices' potential reward. For example, to assess the potential reward of the blue rocket, the model will update its context with the blue rocket and query its episodic memory. It will retrieve a memory that has a mixture of the red and purple planets, though favoring the purple planet, which is the blue rocket's common transition (Fig. 4). It will update its context again with (mostly) the purple planet and retrieve a mixture of (mostly) the two purple aliens residing there and their associated rewards. The memories will be aggregated to arrive at some predicted reward ultimately associated with the blue rocket. The model will then repeat this process with the green rocket and make its choice accordingly.

To match a plausible human choice process, we allowed the model to make a softmax choice, as in Daw et al., 2011, where we replaced the Q-values associated with the candidate choices with the rewards retrieved from memory for each candidate choice,

$$P(c_{i,t} = c) = \frac{\exp(\beta_i[\text{rew}(c) + p \cdot \text{rep(c)}])}{\sum_{c'} exp(\beta_i[\text{rew}(c') + p \cdot \text{rep(c')}])} \quad (2)$$

where we determine the probability of a candidate choice $c_{i,t} = a$ for the rocket or alien stage i during trial t, given the retrieved reward memory of that candidate choice $\text{rew}(c)$. We include a free parameter β_i to titrate how deterministic choices are, and a free parameter p to capture perseverative behavior (Lau & Glimcher, 2005) coupled with the binary function $\text{rep}(a)$: 1 if the candidate choice matches the previous trial and 0 otherwise (only used for rocket choice).

Temporal myopia during memory retrieval captures human & RL models' behavior To reproduce human and RL model behavior in the classic 2-step task, we manipulated the extent to which retrieved memories were temporally myopic (or recency-biased)—while keeping all other parameters constant—by adjusting the size of the weight associated with the timestamp field during memory retrieval ($\alpha_{timestamp}$, Fig. 2). An $\alpha_{timestamp} = 0$ means that the recency of memories does not influence their retrieval, and all memories, old and new, that also match the other queries (state and context) are equally considered. As $\alpha_{timestamp}$ increases, the memories retrieved are biased towards recent memories, resulting in the limit in only the most recent memory being retrieved.

We reproduced human behavior with an intermediate temporal myopia, and MB and MF behavior with low and high temporal myopia respectively (Fig. 5b). Low temporal myopia ("photographic memory") allows the EGO model to leverage all matching memories, old and new, allowing it to best

[2] As in Daw et al., 2011, the EGO model's "choices" (read: preferences) were "off-policy": "choices" on each trial were predicated on the data seen so far, but not implemented because the actual choices were either human or pregenerated.

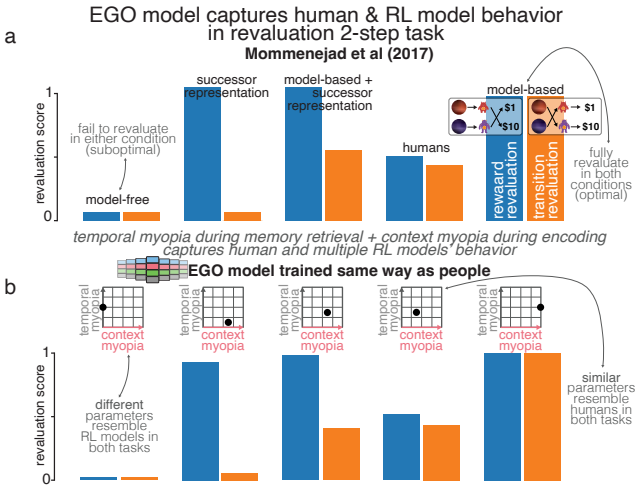

Figure 6: **EGO model captures human & RL model behavior in revaluation 2-step task. (a)** Cartoon of key results from Momennejad et al. (2017): humans revaluate in both conditions, but slightly more in reward revaluation condition, qualitatively resembling a mixture of the successor representation (SR) and MB behavior. **(b)** EGO model trained in the same way as participants: recency-bias during memory retrieval + context myopia (near-sightedness: the extent to which context resembles only the most recent state) can capture human and multiple RL models' behavior. The same parameters resemble human behavior in both tasks, but different parameters are required to resemble RL models' behavior.

reconstruct the task statistics (i.e., the "model" in "model-based"), such as the stable common/rare transitions. High temporal myopia ("goldfish memory") forces the model to only retrieve the most recent memories, it will not have access to the task statistics implicitly stored in memory and instead must rely on its most recent rewards. An intermediate temporal myopia allows some "forgetting" of older memories.

Model trained with human data matches human behavior We initially trained the EGO model with unbiased synthetic data to ensure we were not introducing any biases into the model's memory that could inadvertently aid its performance. It remained an open question, however, whether it required an unbiased sample of the task in order to function as it did. To address this potential issue, we trained the EGO model with 1413 subjects from Gillan et al., 2016. We were able to capture the signature qualities of human behavior in this task (both main effects for transition type and reward outcome) with the exact parameters used to match human behavior when trained with synthetic data (Fig. 5d).

EGO model captures human & RL model behavior in revaluation 2-step task

Revaluation 2-step task The revaluation 2-step task (see Fig. 4 for task description) was designed to test whether rather than be MB (generate an action value from scratch) or MF (look up the precomputed value of an action), people instead did something in between: store partially computed action values (e.g. "my current state's value could be X or Y depending on what happens next") by keeping track of rewards and state transitions separately, a model called the successor representation (SR, Momennejad et al., 2017). In the task, participants are evaluated on their ability to revaluate their starting state preferences based on two conditions, a reward revaluation (unchanged state transitions, but flipped rewards), and a transition revaluation (flipped state transitions but unchanged rewards). Importantly, during the revaluation phases, the sequences are truncated and do not include the starting state, differentiating model behaviors (Fig. 6a):

A MF model fails to revaluate in both conditions because it must see the starting states in order to update their values (i.e., rockets). By contrast, a MB model will equally (and optimally) revaluate regardless of whether it is a reward or a transition revaluation because it can learn and reconstruct the new sequences from scratch. An SR model, which keeps track of rewards and transitions separately, can revaluate in the reward revaluation condition because although incomplete, it can recycle its previously learned state transitions, while separately relearning the new reward contingencies. It cannot, however, revaluate in the transition revaluation condition because the full sequence is not shown, and as such it cannot update its previously learned state transitions beginning with the starting states. Humans are able to revaluate in both conditions (albeit intermediately), and tend to revaluate better in the reward revaluation, interpreted as evidence that humans implement the SR (Momennejad et al., 2017).

Model training and choice process In this task, participants observe randomly interleaved sequences of states during the learning and revaluation phases before making their choices (Fig. 4). We trained the EGO model with the same number of state sequences and simulated participants as in Momennejad et al. (2017). Instead of discrete choices, we extracted the model's internal reward estimates for each starting state, computed the difference between starting states for the learning and revaluation phases, and the difference of differences across the phases as the model's revaluation score.

Temporal myopia during memory retrieval + context myopia during encoding captures human & RL models' behavior To reproduce human and RL model behavior in the revaluation 2-step task, we manipulated the temporal myopia in memory retrieval (as above) as well as how myopic context is during memory encoding. Recall that context (or "working memory") is a linear integrator of the state, and as such, can capture the history of recent states (Fig. 1). If the state integration rate $\gamma = 1$, context will become a carbon copy of state one time step behind, and thus will be fully myopic (Eqn. 1). By contrast, as $\gamma \to 0$, context will reflect a longer sequence of previously seen states, making it less myopic.

We reproduced human behavior in the revaluation 2-step task with similar parameters (intermediate temporal myopia and context myopia) to those in the classic 2-step task (Fig.

6b). The parameters required to capture the different RL models, however, differed between the tasks. MB—i.e., normative—behavior required an intermediate temporal myopia and highly myopic context (vs. low temporal myopia and high context myopia in the classic 2-step task). This is because (1) a photographic memory creates interference in the revaluation task, and (2) because the sequences are not probabilistic, a fully myopic context where only the previous state is recorded in context, will generate the most faithful reconstruction of the task upon memory retrieval. By contrast, MF behavior was replicated by ensuring context was not myopic, creating interference during memory retrieval, even when allowed to have some temporal myopia to reduce that interference from older memories. The SR was replicated with a low temporal myopia and intermediate context myopia (MB in the classic 2-step), and as temporal myopia was increased resembled a mixture of MB and SR.

Discussion

We studied how the EGO framework, a process-model equipped with an episodic memory, a working memory, and a controller (decision-making mechanism), performed on two different RL tasks. Previous work showed that EGO qualitatively resembled human behavior in the revaluation 2-step task (Giallanza et al., 2024). We replicated and expanded on this work by showing that rather than rely on task-specific tuning, the EGO model may capture a task-general process because it accounts for human and RL model behavior in both tasks.

As opposed to typical MF and MB models where values are either explicitly linked to candidate choices, or computed based on an explicit world model, even one that may evolve through a process like latent cause inference (Gershman & Niv, 2012; Gershman et al., 2015), an important quality of the EGO model is that relationships between states and their associated values are *implicit* in its episodic memory, and thus shaped by the encoding and retrieval processes. By titrating encoding and retrieval parameters, we captured the spectrum of human and RL model behavior. Further, we dissociated "semantic time" (context myopia) from actual time (temporal myopia) and showed how they separately contribute to differing task demands. This is an important distinction from other episodic memory models that conflate these through processes like decay on the context vector (Howard and Kahana, 2002).

We reproduced human behavior in both the classic and the revaluation 2-step tasks with similar model parameters. We were also able to mimic RL model behavior in both tasks, but the parameters that matched the same RL models across tasks were markedly different. These results suggest that the underlying process through which people perform these tasks is the same. The normative solutions, however, are fundamentally different across tasks. In the revaluation 2-step task, it is optimal to *forget* the initial sequences in order to properly learn the revaluated sequences. In the classic 2-step task, it is optimal to *remember* most trials in order to properly learn the stable transitions between states. These optimal solutions are, however, highly dependent on the task specifics. The classic 2-step task pairs stable rocket transition probabilities with drifting alien reward probabilities, creating a tension between stable and unstable events. With slowly drifting reward probabilities, as typically formulated, it is optimal to remember well but not perfectly, in order to learn and exploit the stable transition probabilities but not fail to keep up with the drifting reward probabilities. Increase the drift, however, and you must remember less well to remain optimal. Likewise, in the revaluation task, 20 revaluation trials require some forgetfulness, but fewer put a premium on recency. In both tasks, adding a shared "bottleneck" state requires retaining longer sequences in memory (less context myopia), a feature that otherwise creates interference.

Rather than from an inability to behave optimally, people's seeming departure from normative (i.e. model-based) behavior may be due to the fact that (1) an underlying process constrains our behavior, and (2) the parameters for that process reflect the distribution of tasks we face. People may generate "models" on-the-fly by shaping how memories are encoded and retrieved. Performing a short psychological study may simply not be enough time (or reason) to reshape and optimize this process for one specific task, so the strategies implemented may reflect a global optimum. Future work could include modeling other types of RL tasks to investigate whether there is an "optimal" parameter space for real-world choices, and whether there is a metacognitive controller that can optimize memory encoding and retrieval for specific tasks for which it is worthwhile. Behavioral interventions could elucidate these putative processes, e.g. time pressure could increase temporal myopia because of a fixation on time, while higher cognitive load could lead to noise while integrating state into context. Allowing the model to make its own choices (rather than be trained with synthetic or human choices) will allow making predictions as outlined above, as well as exploring the effects of other model elements (choice parameters, time drift, similarity metrics).

Further work could expand the EGO model to other normative cognitive models, like the expected value of control theory (Shenhav et al., 2013). Research has shown that predictions of future learning shape control allocation (Carrasco-Davis et al., 2023; Masís et al., 2021, 2025). These predictions may be generated using a memory retrieval process, rather than a normative but expensive internal simulation (Njaradi et al., 2026). A process-based cognitive modeling approach may be fruitful to understand behavior in tandem with the classic normative-based approach, providing the scaffold on which to compare human and normative behavior, illuminating not only *how* but *why* we solve cognitive problems the way we do.

Acknowledgments We thank Nathaniel Daw for data (Gillan et al., 2016; Momennejad et al., 2017) and technical guidance. JM: NIH 5T32MH065214 training grant and the Latent Cause Inference Conte Center under NIMH P50MH136296. JDC: Vannevar Bush Faculty Fellowship administered by the Office of Naval Research.

References

Anderson, J. R. (1993). *The rules of the mind*. Erlbaum.

Carrasco-Davis, R., Masís, J., & Saxe, A. M. (2023). Meta-learning strategies through value maximization in neural networks. *arXiv preprint arXiv:2310.19919*.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & behavior*, *40*(3), 255–268.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50.

Giallanza, T., Campbell, D., & Cohen, J. D. (2024). Toward the emergence of intelligent control: Episodic generalization and optimization. *Open Mind*, *8*, 688–722.

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

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of mathematical psychology*, *46*(3), 269–299.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, *84*(3), 555–579.

Lebiere, C., & Anderson, J. R. (1993). A connectinonist implementation of the act-r productino system. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *15*.

Masís, J., Musslick, S., & Cohen, J. D. (2021). The value of learning and cognitive control allocation. *Proceedings of the annual meeting of the cognitive science society*.

Masís, J., Musslick, S., & Cohen, J. D. (2025). Learning expectations shape cognitive control allocation. *Proceedings of the National Academy of Sciences*, *122*(44), e2416720122.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature human behaviour*, *1*(9), 680–692.

Njaradi, V., Carrasco-Davis, R., Latham, P. E., & Saxe, A. (2026). Optimal learning rate schedule for balancing effort and performance. *arXiv preprint arXiv:2601.07830*.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Sutton, R. S., & Barto. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge.