In search of lost memories: Modeling exploration with forgetful generalization

Alexandr Ten[1,+,*], **Michiko Sakaki**[1], **Sebastian Breit**[2], **Aswath Chandrasekaran**[2], **Kou Murayama**[1], **and Charley M. Wu**[2,3,4,5]

[1]Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany
[2]Human and Machine Cognition Lab, University of Tübingen, Tübingen, DE
[3]Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, DE
[4]Centre for Cognitive Science, Institute of Psychology, Technical University of Darmstadt, Darmstadt, Germany
[5]Hessian.AI, Darmstadt, Germany
[*]corresponding author, alexandr.ten@uni-tuebingen.de

ABSTRACT

The clarity of our memories guides how we act on past experiences and approach new ones. While extensive research has examined how memory limitations affect recall, far less is known about how these limitations influence decisions in new situations, which depend on generalizing from incomplete or distorted memories. We introduce a computational model of "forgetful generalization", integrating similarity-based generalization with variable-precision memory modulated by recency and asymmetric surprise. In a preregistered spatially correlated bandit experiment, we manipulated (within-subject) whether past observations remained visible (low load) or disappeared (high load). Greater memory load increased "forgetfulness", as captured by parameters indexing recency- and surprise-dependent decay of memory precision. These parameters also predicted individual working memory capacity and explained differences in performance and search patterns. By formalizing how generalization operates under limited memory, our model links episodic reinforcement learning to theories of adaptive memory compression and resource rationality.

Introduction

Imagine showing a visiting friend around your hometown. Some places you can picture in sharp detail, such as the cafe you visited just last weekend. Other memories, such as the park you once loved as a child, have been eroded by time, only allowing partial access to details about what made it special. Then there's the abandoned amusement park you once stumbled upon as a teenager – an unexpected find that still stands out vividly in your memory. Thus, our memories and how clearly we retain them, play an important role in shaping which experiences we revisit, which we avoid, and how we explore new possibilities.

Decades of research have studied how memory limitations shape the accuracy of our recall of past experiences[1–4]. Broad evidence shows how human memory prioritizes certain experiences[5–8], especially more recent[9–12] and surprising events[13,14]. However, we know far less about how memory limitations influence future decisions, particularly when generalizing from incomplete or fuzzy memories to guide choices in new situations.

One extension of reinforcement learning (RL) called episodic RL[15,16] has been central to modeling how past experiences guide future decisions through episodic memory and value generalization[17–19] (Fig. 1a). A prominent example of episodic RL uses the formalism of Bayesian function learning, called Gaussian Process (GP) regression[20], to explain how people infer the value of novel options based on similarities to past experiences[17]. The framework has been successful in accounting for human generalization across spatial[21,22], abstract[23], and graph-structured domains[24]. Episodic RL, however, assumes that past experiences are stored with perfect precision, overlooking how cognitive limitations reduce representational fidelity[25] (e.g., over time), and in turn, constrain generalization[17]. Therefore, it cannot account for how memory distortions influence generalization, thus altering decision-making and exploration.

Another extension to RL provides a mechanistic account of how working memory (WM) limitations influence learning[26–28] (Fig. 1b). These RLWM models combine fast, accurate WM encoding with slower RL updates. When WM capacity is sufficient to meet task demands, reward associations can be learned in a single trial; under higher load, learning proceeds more gradually via RL (also see REF[29]). Crucially, RLWM models provide a theoretical link between episodic and working memory, offering a mechanistic account of how both the encoding and recall of episodic information can be impaired by the lack of attentional resources, resulting in less precise memories[30,31].

Related work on exploration has examined how a broad range of cognitive constraints, including both WM load[32] and time pressure[33,34], affect the balance between exploiting known rewards and exploring uncertain options. Across studies, these constraints selectively reduce *uncertainty-directed exploration*, while leaving *random exploration* relatively unaffected[35,36]. Together, these lines of research show how memory and cognitive constraints degrade the precision of value representations and limit the strategic use of uncertainty for exploration. Yet, much like RLWM models, they share the

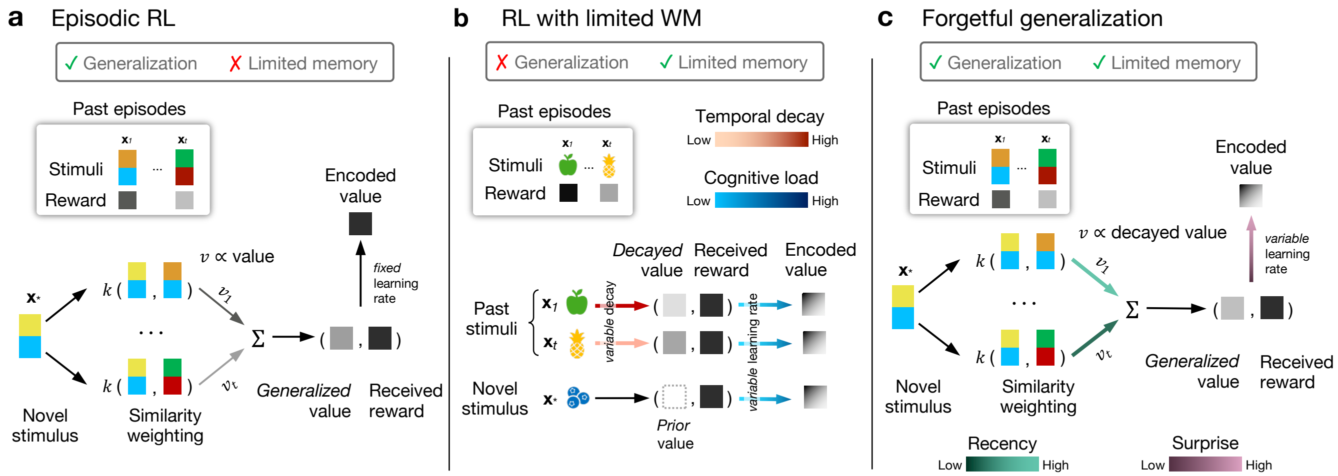

Figure 1. Theoretical frameworks. **a**) Episodic reinforcement learning (RL) makes predictive generalizations about the expected value of a novel stimulus $\mathbf{x}_*$ by computing a weighted sum of similarity scores (e.g., using kernel similarity; Eq. 3) between $\mathbf{x}_*$ and episodic memories of previously encountered stimuli, which are assumed to be stored with perfect precision. **b**) RL with limited working memory (RLWM) accounts for how cognitive load (e.g., high learning rate under minimal load) and temporally decay influences value learning, which is learned independently for all stimuli (i.e., no generalization). **c**) In our framework, value is generalized from the past episodes (as in panel **a**), but episodes are stored with variable-precision, causing both temporal- and surprise-dependent decay (as in panel **b**).

same key simplifying assumption: each option is independent, such that knowledge about one provides no information about others. This overlooks the hidden structure of real-world environments[21,37], which allow for generalization from familiar to novel situations.

Here, we introduce a new model of *forgetful generalization* (Fig. 1c), which integrates the mechanisms of value generalization[17] with principles of RLWM[28] under the statistical framework of heteroskedastic GP regression[20,38,39]. As in the previous work using GPs as a Bayesian framework for value generalization (and episodic RL more broadly) value is inferred on the basis of similarity to past episodes. However, in our model, past episode are encoded with *variable precision*[40], capturing loss of fidelity due to memory limitations. This allows us to explore mechanistic hypotheses about how the encoding of events are prioritized (e.g., through recency and surprise), and how memory distortions influence decision-making and exploration.

To link memory prioritization to individual differences in WM, we assessed each participant's WM capacity using the symmetry span task[41,42] (Fig. 2a). We then evaluated the forgetful generalization model (and lesioned variants) in predicting human behavior in a variant of the spatially correlated bandit task[21] (Fig. 2b). The task presents participants with 121 different options on a 11×11 grid with hidden spatial structure, such that nearby options have similar rewards. This large and structured decision space contains far more possible options than can be sampled within the available search horizon (25 trials), making generalization and efficient exploration crucial for good performance. To understand how memory limitations influence these capacities, we additionally manipulate memory load (within-subject). In the Low Load condition (LL), reward observations remain visible, whereas in the High Load (HL) condition, they disappeared after 400 ms requiring participants to rely on memory to generalize about the expected value of novel options.

Results

We collected data from $N = 197$ participants to test how working memory (WM) capacity and cognitive load influence generalization and exploration in a spatially correlated bandit task (Fig. 2; see Methods). We first report behavioral effects before turning to a new computational model of forgetful generalization, which provide a richer, mechanistic account. The study design, exclusion criteria, hypotheses, and analysis plan were preregistered prior to data collection[43].

Behavioral results

Before introducing the computational model, we examined how memory load (low load, LL; high load, HL) and individual WM capacity (symmetry span task) influenced both performance and search patterns on the bandit task. WM scores were computed using partial-credit scoring, which takes the mean proportion of correct responses across all sequences[44]. Welch's t-test indicated that our sample ($M = 0.69$, $SD = 0.21$) did not significantly differ from a normative sample[42] ($t_{222.1} = 1.44,\ p = .150$).

Bandit performance

We preregistered the prediction that the high load (HL) condition would impair performance relative to the low load (LL) condition. A paired t-test on mean normalized rewards did not support this prediction ($t_{196} = 1.35,\ p = .179$; see Methods for

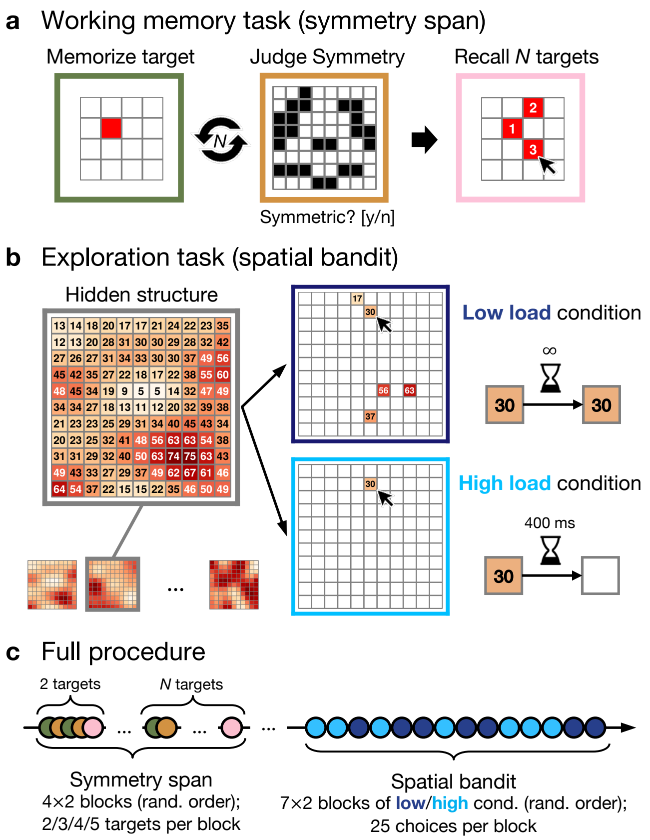

Figure 2. **a**) Symmetry span task used to assess memory capacity. Participants alternated between memorizing target locations presented as red squares and judging whether black-and-white patterns were symmetric along the vertical centerline. After completing the sequence, participants sequentially reproduced the target locations on an empty grid. **b**) Spatially correlated bandit task with variable cognitive load. Each 121-armed bandit was represented as an 11×11 grid with a hidden spatial structure, where nearby arms had similar values. In the "Low load" (LL) condition, payouts remained visible until the end of the block. In the "High load" (HL) condition, payout information disappeared after 400 ms, requiring participants to rely on memory. **c**), The full procedure consisted of the symmetry span memory task followed by the bandit task, where memory load was manipulated within-subject and appeared in randomized order.

details on reward normalization). However, exploratory analyses revealed that memory load affected how participants accumulated rewards over time (Fig. 3a). For this analysis, we fit a linear mixed-effects regression on the difference in mean normalized reward between conditions (LL - HL) as a function of trial number: `diff ~ trial + (trial|id)`. A positive intercept ($b = 0.018,\ t_{196} = 2.32,\ p = .022$) indicated participants reliably obtained higher rewards on their first free choice (i.e., the second trial) in LL compared to HL, while a negative trial slope ($b = -0.022,\ t(196) = -2.08,\ p = .038$) suggested this advantage diminished over the course of the block (Fig. 3a, inset).

We next explored whether WM scores predicted performance. A preregistered correlation between WM score and the LL-HL performance difference was not significant ($r_{195} = .035,\ p = .625$), providing no empirical support for the hypothesis that WM capacity moderated the effect of memory load on overall performance. However, an exploratory mixed-effects model of mean performance as a function of condition (effects-coded) and WM score (mean centered), `reward ~ cond * WMScore + (cond|id)`, revealed a significant relationship between WM score and overall performance ($b = .024,\ t_{195} = 3.69,\ p < .001$; Fig. 3b; see Fig. S1 for all model coefficients). Thus, participants with a higher WM score performed better on the bandit task overall, regardless of load conditions.

Search distance

We then explored differences in search distance between the LL and HL conditions (Fig. 3c-d), which we defined as the Manhattan distance between consecutive choices. A t-test showed participants searched more locally under HL ($t_{196} = 4.30,\ p < .001$). Stratifying by trial and using a linear mixed-effects regression, we found that this difference was stable across the block ($b = -0.022,\ t_{196} = -0.19,\ p = .852$; Fig. 3c).

To further characterize the differences in search patterns, we classified participants' choices into 'repeat' (dist = 0), 'near' (dist = 1), and 'far' (dist > 1) bins and examined how memory load changed the frequency of these choices (Fig. 3d). Although the number of repeat choices did not differ significantly between conditions ($b = .344,\ t_{196} = 1.78,\ p = .077$), the frequency of near choices was higher in the HL condition ($b = .935,\ t_{196} = 5.62,\ p < .001$), while the frequency of far choices was higher in the LL condition ($b = -1.279,\ t_{196} = -6.59,\ p < .001$). Thus, differences in search distance were primarily driven by a change in the nature of exploration (less distant under HL), rather than a change in the balance between exploration vs. exploitation (similar repeat choices).

Reward sensitivity

We next examined whether memory load influenced reward-guided search. Previous studies[22,39] found higher rewards at time t lead to shorter search distances at $t+1$ (reward sensitivity), which can be considered a flexible analog of a simple win-stay-lose-shift heuristic[45]. If memory load disrupts value representations, reward sensitivity could weaken under HL, which could be evidenced by a significant interaction between reward and condition (in the presence of a significant effect of reward).

Our preregistered mixed-effects Poisson regression, `dist ~ reward * cond + (reward * cond|id)`, confirmed a robust reward sensitivity in LL ($b = -2.68,\ z = -30.89,\ p =< .001$) and shorter overall search distances in HL ($b = -0.124,\ z = -6.98,\ p =< .001$), but did not detect a significant reward $*$ condition interaction ($b = -0.066,\ z = -1.50,\ p = .134$). However, the effect estimate was in the

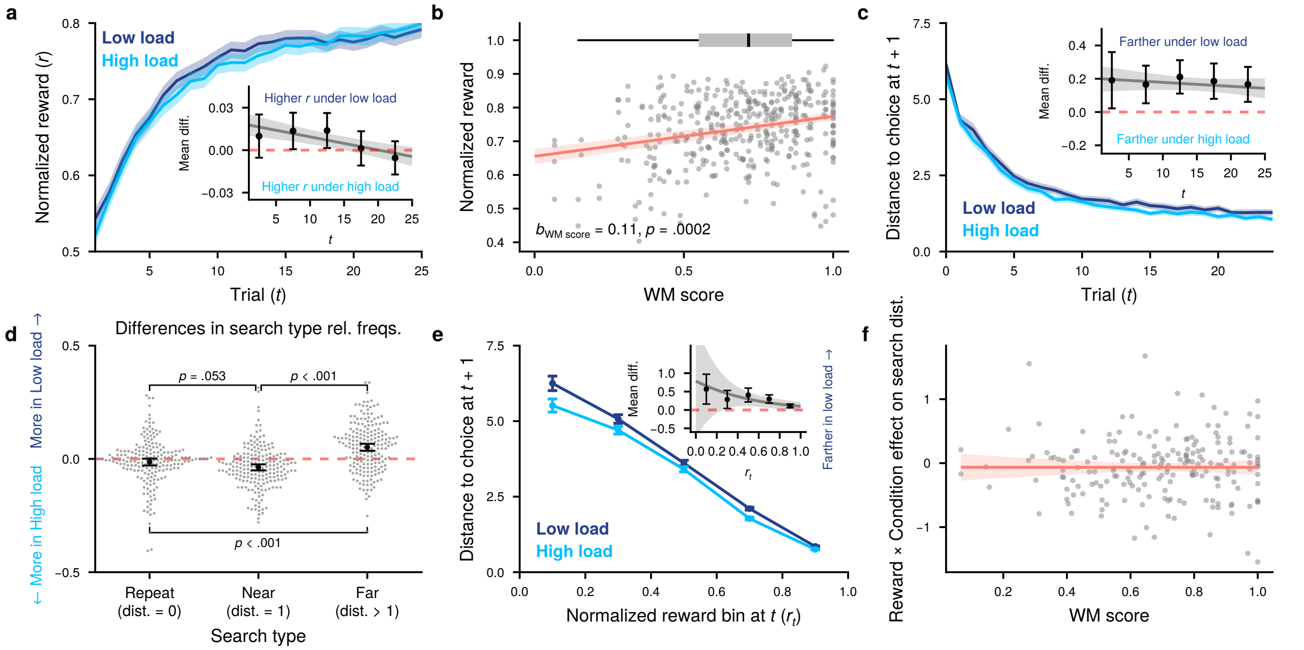

Figure 3. Behavioral results. **a**) Mean normalized reward (±95% CI) across participants and blocks, stratified by trial (x axis) and condition (colors). The inset shows binned, within-participants mean differences between conditions (with 95% CI), where the gray line indicates the expected value of differences (with 95% CI) as a function of trial. **b**) Joint distribution of mean normalized rewards and individual WM scores. The data points are colored according to 33%-percentiles designating low, medium, and high WM scoreres. The box plot on top shows the interquartile range and the median of WM scores. The red line shows the expected value (with 95% CI) of mean normalized reward as a function of WM score. **c**) Search distance (±95% CI) across participants and blocks, stratified by trial (x axis) and condition (colors). The inset shows binned, within-participants mean differences between conditions (with 95% CI), where the gray line indicates the expected value of differences (with 95% CI) as a function of trial. **d**) Distributions of differences in the relative frequencies of choices of repeat, near, and far choices. Points above the red dashed line indicate higher frequency in the Low load condition. Black dots and with error bars in the middle of each distribution indicate the mean and the 95% CI. Labeled brackets connecting pairs of distributions summarize the corresponding t-tests. **e**) Search distance averaged across participants and blocks (±95% CI), stratified by binned reward (x axis) and condition (colors). The inset shows the average difference in search distance for each reward bin (with 95% CI), along with the predicted differences (±95% CI) between conditions predicted by the preregistered Poisson model described in Results. **f**) Joint distribution of WM scores and estimated random effects of participant on Reward $*$ Condition interaction (line and band represent the expected value ±95% CI of random effect conditional on WM score). WM score did not seem to modulate the effect of condition on reward sensitivity.

expected direction and the average (within-participant) HL-LL contrast in distance appeared to decrease as a function of reward (Fig. 3e, inset). Thus, we investigated this further in an exploratory analysis using a *linear* mixed-effects regression (as used previously[22,39]), which did in fact indicate a significant interaction ($b = .610,\ t_{192.48} = 4.91,\ p =< .001$). However, model comparisons proved inconclusive (Fig. S2), thus we cannot definitively confirm or reject the modulatory effect of memory load on reward sensitivity.

Finally, we tested the hypothesis that lower WM capacity accentuated memory load impairments of reward-guided search. The preregistered mixed-effects Poisson model, `dist ~ reward * cond * WMScore + (reward * cond|id)` found no three-way interaction ($b = -0.188,\ z = -0.90,\ p = .366$), indicating no moderation of the load effect by WM score (Fig. 3f, see Fig. S3 for all model coefficients and tests). Thus, the effect of memory load (or lack thereof) did not appear to depend on individual WM capacity.

Behavioral summary

Memory load produced subtle but consistent effects: HL reduced early trial rewards, shortened search distances through increased number of near choices (dist. $= 1$) and reduced number of far choices (dist. > 1). While one might expect decreased search distance could be due to weaker reward sensitivity under HL, our data did not support this prediction. Furthermore, while individual WM capacity boosted overall performance, it was not predictive, in either condition, of search distances across different reward levels. These patterns hint at interactions between memory constraints and generalization, but behavioral analyses alone cannot reveal the underlying mechanisms. Next, we directly assess these mechanisms using our computational model of forgetful generalization.

Computational modeling

We introduce a computational model of "forgetful generalization" (along with lesioned variants; Fig. 4a, b) to provide an algorithmic account of how memory constraints shape generalization and decision-making, thus unifying mechanisms of value generalization from episodic RL (Fig. 1a) with the memory dynamics of RLWM (Fig. 1b). In contrast to prior work using Gaussian Process (GP) regression, and the episodic RL framework more generally, our model assumes past episodes are stored with *variable precision*, reflecting the limited and selective nature of memory (Fig. 1c).

Specifically, we test mechanisms prioritizing the precision of past experiences according to their *recency*, the degree of signed *surprise* given prior expectations, and their interactions (Fig. 4c). These mechanisms of forgetful generalization extend the standard GP-UCB framework[17], which combines GP regression for value generalization and upper-confidence-bound (UCB) sampling for exploration.

Value Generalization

GP value generalization is performed using Bayesian inference about the expected rewards for a target location on the grid $\mathbf{x}_\star$, conditioned on past observations $\mathscr{D}_t = \{\mathbf{X}_t, \mathbf{r}_t\}$ of choices $\mathbf{X}_t = [\mathbf{x}_1, \ldots, \mathbf{x}_t]$ and associated rewards $\mathbf{r}_t = [r_1, \ldots, r_t]$ at time t. In our particular setting, the posterior predictive distribution takes the familiar Gaussian form $p(r(\mathbf{x}_\star)|\mathscr{D}_t) \sim \mathcal{N}(m_t(\mathbf{x}_\star), v_t(\mathbf{x}_\star))$, with mean and variance:

$$m_t(\mathbf{x}_\star \mid \mathscr{D}_t) = \mathbf{k}_\star \left[K_{X,X} + \sigma_n^2 I\right]^{-1} \mathbf{r}_t \tag{1}$$

$$v_t(\mathbf{x}_\star \mid \mathscr{D}_t) = 1 - \mathbf{k}_\star^\top \left[K_{X,X} + \sigma_n^2 I\right]^{-1} \mathbf{k}_\star. \tag{2}$$

Here, $\mathbf{k}_\star = [k(\mathbf{x}_1, \mathbf{x}_\star), \ldots, k(\mathbf{x}_t, \mathbf{x}_\star)]$ is the vector of kernel similarities between past observations and the target location (illustrated in Fig. 1a), and $K_{X,X}$ is a matrix of pairwise kernel similarities between all past observations in $\mathbf{X}_t$. To define similarity, we adopt the common radial basis function (RBF) kernel:

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right), \tag{3}$$

where the lengthscale parameter λ controls the degree to which rewards generalize over space.

An important link to episodic RL[15,46] (Fig. 1a) emerges from the fact the posterior mean (Eq. 1) can be re-expressed[17,47] as a similarity-weighted sum over value (v) of past episodes (see Fig. 1a)

$$m_t(\mathbf{x}_\star \mid \mathscr{D}_t) = \sum_i^t k(\mathbf{x}_i, \mathbf{x}_\star) v_i, \tag{4}$$

where $v_i \in \mathbf{v} = \mathbf{r}_t \cdot [K_{X,X} + \sigma_n^2 \mathbf{I}]^{-1}$. This formulation clarifies the mechanistic assumption that generalization is achieved by retrieving prior episodes and integrating them based on (a) their similarity to the current situation $k(\cdot, \mathbf{x}_\star)$ and (b) how much weight they carry v. This weight depends jointly on the observed rewards and the assumed reliability, with the latter governed by the observation noise parameter σ_n^2. In standard GP regression, σ_n^2 is fixed to a constant, implying that all episodes are remembered with equal precision. In the next section, we relax this assumption by allowing this noise parameter to vary across episodes. This enables us to formulate a mechanistic account of how some memories are stored with higher fidelity than others (Fig. 4c).

Memory prioritization

We operationalize variable precision memory[3,40,48] using the statistical framework of heteroskedastic GP regression[38,39]. Related approaches have used heteroskedastic GP regression in simpler contexts (e.g., assigning different noise levels to individually experienced vs. socially observed rewards[49]). Here, we extend this idea to model how memory precision varies continuously with cognitive factors such as recency and surprise. Intuitively, past observations (i.e., memories) are stored with greater precision when they are associated with less noise σ_n^2. This means that the greater the noise, the more the model posterior decays back to the prior (similar to REF[26–28]). We define σ_n^2 for each observation as a function of fixed baseline noise $\sigma_0^2 = 10^{-4}$ plus an observation specific variance gain $g_t(\mathbf{x}) \geq 0$:

$$\sigma_n^2 = \sigma_0^2 + g_t(\mathbf{x}) \quad \text{where} \quad g_t(\mathbf{x}) = \exp(\mathbf{f} \cdot \mathbf{w}) - 1 \tag{5}$$

Thus, variance gain is an exponentiated linear combination of features $f_i \in [0, 1]$ and weights $w_i \geq 0$. Features capture various mnemonic prioritization mechanisms (Fig. 4b), such as recency[11,13,50] and surprise[13,14], while weights are free parameters estimated to determine their contribution such that $g_t(\mathbf{x}) = 0$ when all weights are zero.

Recency is defined as the number of elapsed trials since observation $\mathbf{x}$, normalized by the maximum number of trials $t_{\max} = 25$:

$$f_{\text{r}} = \delta(\mathbf{x}, t)/t_{\max}. \tag{6}$$

The higher the corresponding weight w_{r}, the greater the noise associated with older observations (i.e., memory decay). Note that f_{r} is actually the inverse of recency, but this intentional formulation allows us to simplify the interpretation of the weights in $\mathbf{w}$.

Surprise is defined by the mismatch[51] at time t between the observed reward $r_t(\mathbf{x})$ and the predictive mean $m_t(\mathbf{x})$ (i.e., the

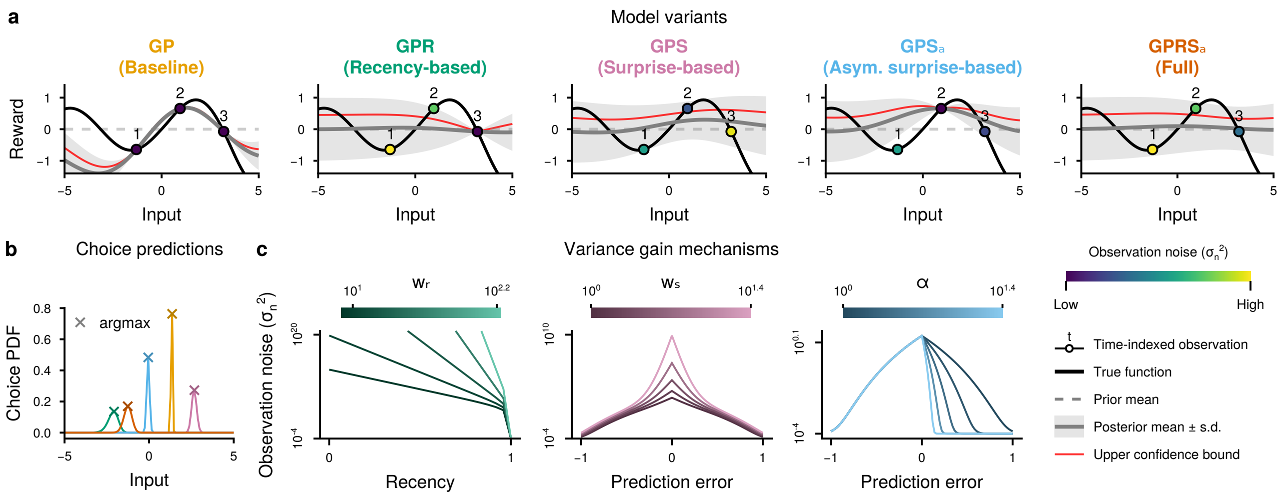

Figure 4. Computational modeling. a) Illustrative comparison of different computational models in a simplified (1D) setting with a continuous input. Each panel shows model predictions after observing the same three rewards. The rewards are given by the true reward function (black curve). The numbers above observations indicate temporal order. The colors of observations indicate observation noise (σ_n^2). The gray line indicates posterior predictive mean conditioned on all 3 observations, and the surrounding band corresponds to posterior predictive variance. The red line signifies the upper confidence bound (UCB), under a constant directed-exploration (β) parameter. **b**) Probability of choosing different inputs given model predictions in **a**. **c**) Variance gain mechanisms. Generally, recency and surprise decrease variance gain (and thus, observation noise). The effect of recency and surprise (a monotonic function of prediction error) is controlled by the corresponding prioritization weight (w_r or w_s), such that greater weights correspond to more extreme prioritization. Additionally, the asymmetry parameter α, controls the asymmetry in how negative vs positive prediction error affect variance gain.

expected reward at time t):

$$f_\mathrm{s} = 1 - \mathrm{erf}\left(\frac{|m_t(\mathbf{x}) - r_t(\mathbf{x})|}{\sqrt{2v_t(\mathbf{x})}}\right), \tag{7}$$

where erf is the Gauss error function, which in our case, is the cumulative density function of a half-normal distribution with standard deviation defined by the prior predictive variance $\sqrt{v_t(\mathbf{x})}$. The farther an observation $r_t(\mathbf{x})$ is from the expectation $m_t(\mathbf{x})$, the closer the error is to 1. The corresponding weight w_s defines the degree with which less surprising observations are forgotten.

We also include an *asymmetric* variant of surprise-based forgetting[13,14], which differentiates positive and negative prediction errors with a surprise-asymmetry parameter α:

$$f'_\mathrm{s} = \begin{cases} \alpha \cdot f_\mathrm{s} & \text{if } r_t(\mathbf{x}) > m_t(\mathbf{x}), \\ f_\mathrm{s} & \text{otherwise.} \end{cases} \tag{8}$$

Thus, if the reward is greater than expected, the surprise feature f_s is multiplied by α. When $\alpha > 1$, positive prediction errors have greater mnemonic priority compared to negative prediction errors, and vice versa for $\alpha < 1$.

Exploration and choice

Following the GP-UCB framework[17], we model choices using a softmax function over the upper confidence bound (UCB) values associated with each option. UCB values are computed as a linear sum of the posterior predictive mean and variance: $\mathrm{UCB}(\mathbf{x}) = m(\mathbf{x}|\mathscr{D}_t) + \beta\sqrt{v(\mathbf{x}|\mathscr{D}_t)}$, where β is the exploration bonus, defining how much the reduction of uncertainty is traded-off against exploiting immediate rewards. The softmax choice rule, $p(\mathbf{x}) \propto \exp(\mathrm{UCB}(\mathbf{x})/\tau)$, includes the temperature parameter $\tau > 0$, which defines the degree of random exploration.

Together with a "baseline" model without forgetting, we consider a set of five models (Fig. 4a) with different combinations of recency (Eq. 6), surprise (Eq. 7), and asymmetric surprise (Eq. 8). As illustrated in Figure 4b, the models can make distinct predictions given the same set of observations.

Model results

Model estimation was based on leave-one-block-out cross-validation (CV) using a differential evolution optimizer[52]. We fitted each model to each condition of each participant. We then performed hierarchical Bayesian model selection[53] to infer the most likely model in the population (protected exceedance probability; PXP).

In both conditions, the full model with recency and asymmetric-surprise forgetting (GPRS$_\mathrm{a}$) had the highest PXP (Fig. 5a) and reliably lower CV loss compared to all other models (Fig. 5b). Crucially, all models were highly recoverable (see S4), allowing us to reliably identify the true generating

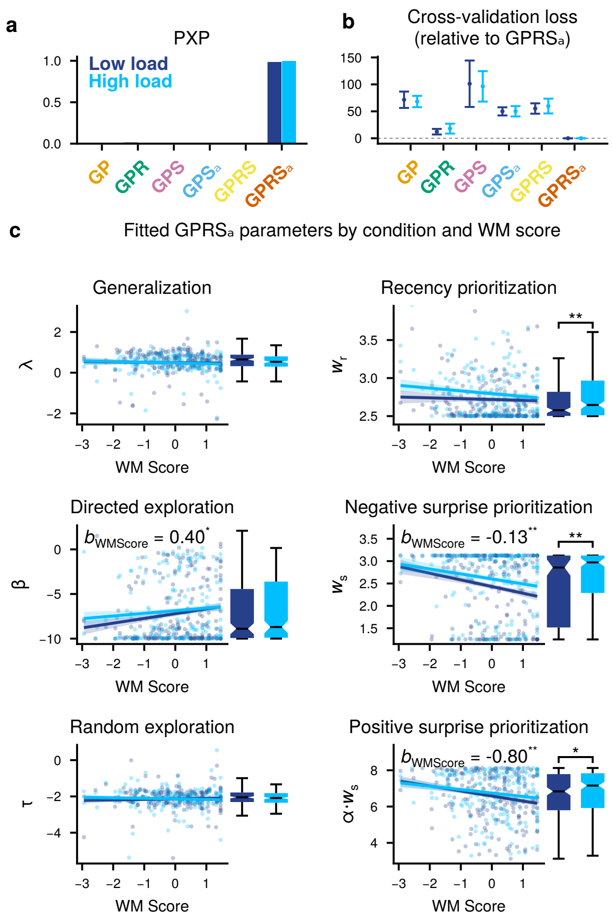

Figure 5. Model results. **a**) Model comparison using protected exceedance probability (PXP). The bars indicate the probability that a given model (x-axis) is most likely in a population (separate for LL and HL), while correcting for chance. **b**) Mean cross-validation loss (±95% CI) relative to the winning model ($GPRS_a$), where larger values are worse. **c**) Joint distributions of the full $GPRS_a$ model's parameters (fitted separately to Low and High load data) and individual WM scores. Note the different scales on the y-axis. Box plots depict marginal distributions of the parameters for each group. The lines illustrate linear mixed-effects models' predictions of parameter values as functions of WM score and condition. Brackets above the box plots summarize the results of significance tests on the fixed-effects of condition on fitted parameters, with * indicating $p < .05$ and ** indicating $p < .01$

model when simulating behavior[54]. Interestingly, the "forgetful model" performed best in the LL as well as HL conditions, suggesting that the model also captures attentional limitations account for all observations, even when they do not disappear.

We then examined the parameter estimates of the winning $GPRS_a$ model, which were also recoverable (see Fig. S5), to examine how they varied with cognitive load and WM scores. Specifically, we fit a linear mixed-effects model (Fig. 5c; see Fig. S6 for Regression coefficients and significance tests), modeling each parameter as a function of condition and WM score: `param ~ cond * WMScore + (cond|id)`. While w_s and α were estimated as separate parameters, we can interpret $\alpha \cdot w_s$ as a single parameter controlling the effect of positive surprise, in contrast to w_s for negative surprise. Thus, we entered w_s and $\alpha \cdot w_s$ as separate dependent variable in the regressions.

First, we observed an effect of condition on all three prioritization parameters (w_r, w_s, and $\alpha \cdot w_s$; see box plots in Fig. 5c), with HL predicting greater recency weights ($b = -0.04,\ t_{195} = -3.28,\ p = .001$), and both higher negative ($b = -0.09,\ t_{195} = -3.10,\ p = .002$) and positive surprise weights ($b = -0.33,\ t_{195} = -2.43,\ p = .016$). There were no reliable effects of condition on the generalization λ, directed exploration β, and random exploration τ parameters (all $p > .242$). Thus, HL specifically increased recency and surprised-based prioritization of reward information.

Next, we looked at how individual WM scores influenced parameters in each condition. While neither generalization λ nor random exploration τ were influenced by WM score (all $p > .612$), participants with higher WM scores had greater β estimates ($b = .41,\ t_{195} = 2.08,\ p = .039$). WM scores did not interact with condition ($b = .12,\ t_{195} = .81,\ p = .419$). Thus, greater WM capacity contributed to more directed exploration in both conditions, consistent with past work linking the two[32].

And while WM scores had no significant effect on recency prioritization ($b = -0.02,\ t_{195} = -0.97, p = .332$), higher WM scores predicted a lower degree of both negative (w_s: $b = -0.13,\ t_{195} = -2.99,\ p = .003$) and positive surprise-based prioritization ($\alpha \cdot w_s$: $b = -0.80,\ t_{195} = -3.06,\ p = .003$). Thus, individuals with high WM had less differential prioritization of surprising vs. unsurprising information, whereas those with low WM prioritized surprising observations to a greater degree, commiserate with their greater need to efficiently allocate limited resources.

Fitted model parameters also showed negative effects of prioritization parameters (w_r, w_s, and $\alpha \cdot w_s$) on overall performance and reward sensitivity (see Fig. S7). This suggests that individuals with stronger forgetting of older and unsurprising observations performed worse, and adapted their choices to observations less effectively.

Discussion

The assumption that future choices depend on representations stored in memory is ubiquitous among RL models of learning and exploration. Yet, little is known about how memory limitations influence generalization and exploration in new situations. In this study, we manipulated cognitive load by having participants search for spatially correlated rewards with either permanent (low load; LL) or disappearing reward observations (high load; HL). Under HL, where representations of past outcomes are likely to impaired, participants were less efficient at maximizing rewards and searched more locally. Furthermore, lower working memory (WM) capacity reliably predicted lower average performance. Overall these results align with the assumption that memory limitations have implications for future decisions via value generalization[17].

To formalize the link between memory limitations and generalization, we integrate the episodic RL framework with ideas from RL with limited WM. Whereas standard GP models assume each episode is stored with equal precision, we introduce a boundedly rational account of "forgetful generalization" where experiences are encoded with *variable precision*, reflecting systematic distortions introduced by memory limitations[40]. Specifically, we implement this using the statistical concept of *heteroskedastic noise*, allowing some experiences (e.g., more recent and surprising events) to be represented sharply, while others are degraded. This new framework provides both an intuitive statistical formalization of memory prioritization, but, as confirmed by model and parameter recovery (see Fig. S4-S5), also offers an effective empirical framework to test different mechanistic hypotheses about which factors compete for limited memory resources.

Our main prioritization mechanisms included modulation of observation noise by recency, surprise, and outcome asymmetry (positive vs. negative surprise), which are factors that have empirical links to forgetting[27] and learning[13,14]. The winning GPRS$_a$ model incorporated all three mechanisms and beat all lesioned variants, suggesting that each mechanism uniquely contributes to explaining how memory limitations shape generalization and decision-making. Recency captures the natural decay of memory fidelity over time[9–12], surprise reflects the heightened salience of unexpected outcomes[13,14,40], and asymmetry allows for differential impacts of better- or worse-than-expected outcomes[14].

Both our experimental manipulation of cognitive load (LL vs. HL) and individually measured WM scores were captured by changes in model parameters. Specifically, participants increased their prioritization of recent (w_r) and surprising events (asymmetrically more for positive $\alpha \cdot w_s$ vs. negative outcomes w_s) under higher load (HL), while participants with lower WM scores had generally higher levels of prioritization for these same factors. In addition, WM capacity predicted the degree of directed exploration, with higher WM associated with greater reliance on uncertainty-directed search (β) Thus, when less memory resources are available (by either situational load or individual capacity), they become more effectively prioritized towards recent and surprising events, particularly for positive, better-than-expected outcomes, while also shaping how effectively uncertainty is leveraged for exploration. Taken together, these findings suggest an adaptive allocation of memory resources, at both short timescales (between LL and HL rounds) and at the population level (across individual differences in WM capacity).

We also recognize that the full GPRS$_a$ model provided the best predictions in both LL and HL conditions. This suggests that even when veridical records of past experiences are available (LL condition), value generalization may nevertheless involve imperfect representations and/or integration of data. One likely possibility is that participants did not always attend to or encode all available reward information, especially given the sheer size and complexity of the decision space (i.e., 121 options). Much like ordering at a restaurant without carefully considering every option on the menu, participants may have selectively focused on a subset of observations, effectively "forgetting" or ignoring the others, despite their availability. Given that we had the same winning model with the same prioritization mechanisms in both conditions, we may speculate that similar prioritization mechanisms exist for both attention and memory[6,55]. This could be based on a general principle that when cognitive resources are constrained (be it through memory or attentional factors), they should be reallocated as effectively as possible. However, more research is required to support this hypothesis. Here, we set out to develop a model of "forgetful generalization", which may in fact be more general than intended, capturing both attentional bottlenecks as well as memory limitations.

Formally, our heteroskedastic GP model can be viewed as introducing an adaptive form of Bayesian regularization[40]. By scaling down the influence of older or less surprising observations on prediction, the model behaves more conservatively in how it generalizes (i.e., an analogue of Tikhonov regularization[56,57], where observation-specific noise modulates how closely predictions follow the data vs. the prior). This comes at the cost of precision, as reflected in the negative relationship between prioritization weights and performance (Fig. S7). Yet, noisy representations can also be adaptive[58] in some settings. Here, we specifically observe a resource-rational use of limited resources[8,25,59,60], where participants selectively allocate limited representational capacity to information that is most likely to improve predictions. From this perspective, forgetting is not merely a limitation but an adaptive feature of cognition[61], ensuring scarce resources are effectively directed towards recent and surprising events, which are most informative for updating believes about a structured, yet uncertain environment.

A first limitation concerns how our model conflates working memory and episodic memory. Although it is reasonable to assume that decreased working memory resources (due to load) would reduce the fidelity with which episodic experiences are encoded or maintained[62–64], our model does not explicitly distinguish between the active maintenance of information in working memory and the longer-term storage and retrieval of episodic memory[65]. By treating them as a single source of representational noise, we may blur potentially important mechanistic differences[66,67]. Future work could address this by extending our framework to include separate and distinct contributions both systems, or by combining modeling with neural data to isolate their distinct signatures.

Another limitation lies in the strength of our memory load manipulation. Our behavioral paradigm manipulated the availability of outcome information (persistent in LL and disappearing after 400ms in HL), which revealed reliable effects on performance, exploration patterns, and model-based parameter estimates. However, not all of our preregistered predictions[43] were validated by the data. One potential reason is that the HL condition did not sufficiently tax participants' memory resources. Other manipulations, such as visual-noise masking[68], a parallel WM load task[32], fixed delays between choice and feedback[69], or imposing time pressure[33,70], may be necessary to ensure stronger behavioral effects. Thus, new experimental designs may be necessary to further clarify the effect of memory demand, and to additionally understand how it interacts with other resource constraints, such as attention.

In conclusion, our study provides a first step towards a mechanistic account of how memory limitations distort value generalization and exploration. By integrating memory-prioritization processes with a Bayesian model of episodic RL, we show that both experimental manipulations of load and individual differences in WM capacity systematically modulate the precision of stored experiences and the use of uncertainty-directed exploration. Under high load and for participants with lower WM capacity, prioritization was increased for recent and surprising events, consistent with a resource-rational allocation of limited representational capacity. Together, these findings highlight how forgetting is not only a costly error, but also an adaptive feature of cognition, shaping how people learn from the past to guide decisions in complex environments.

Methods

The study protocol was preregistered on the Open Science Framework platform and carried out accordingly[43]. The experimental procedure was reviewed and approved by the Ethics Committee of the Faculty of Economics and Social Sciences, University of Tübingen. Informed consent was obtained from all participants. Participants were paid a fixed base fee of £4, plus the bonus amount up to 100% of the base fee earned during the tasks.

Participants

We recruited $N = 200$ participants (53.4% male, 46% female, 1 non-binary) from the Prolific US (59.7%) and UK pools. All participants reported fluency in English. We excluded 3 participants who self-reported using external aids (e.g., notes) during the tasks. The age ranged between 19 and 74 years ($M = 39.36$, $SD = 12.95$). Data from three participants were excluded from analyses due self-reported use of external aids (e.g., written notes).

The number of participants was determined by a priori power analyses conducted in `G*Power v3.1`[71]. The power analyses were conducted for tests of the preregistered hypotheses, with the goal of obtaining .8 power to detect a small-to-medium effect size of $d = .20$ (comparison of mean reward levels between conditions) and $r = 0.20$ (correlation between working memory score and the effect of condition) as the smallest effect size of interest at the significance level of .05.

Study design

Participants underwent a three-part ordered procedure (online) consisting of a symmetry span task (Fig. 2a), a spatially correlated bandit task (Fig. 2b), and a demographics and personality questionnaire. During the spatially correlated bandit task, we manipulated memory load (within-subject), such that half of the 14 blocks were randomly assigned to the low load condition (LL), and the other half to the control high load condition (HL), with the conditions presented in randomized, interleaved order. In the LL condition, reward observations were visible until the end of the block, whereas they disappeared after 400 ms in the HL condition, requiring participants to rely on memory to guide their exploration.

Participants could earn a bonus of up to a 100% of the base fee. Half of the bonus points could be earned by maximizing performance in the symmetry span task, and the other half by maximizing performance in the bandit task. The median completion time was ≈ 31 minutes and the participants earned an average of £6 (including the bonus).

The study was conducted online using custom and Labjs[72] code served with JATOS[73]. The procedure consisted of two main parts: the symmetry span memory task and the spatially correlated bandit task. After providing informed consent, participants read the overview of the entire procedure and then continued to the memory task, followed immediately by the bandit task. At the end of the study, we administered a questionnaire collecting demographic data (age, gender, education) and measuring the Need for Cognition trait[74]. Participants read a dedicated set of instructions before each task. Participants could not start the behavioral task unless they correctly answered procedure-comprehension questions.

Symmetry span task

The *symmetry span task* is a validated variant of a widely used complex span paradigm for measuring the ability to maintain target information under interference caused by complex task demands[44]. Our implementation followed an existing shortened procedure[41,42]. Participants completed sequences of sizes between 2 and 5, with each sequence repeated twice (for a total of 8 unique sequences). A single trial required participants to memorize (in correct order) the locations of sequentially presented visual stimuli on a 4×4 grid, with sequences interleaved with independent trials of a distractor "processing" task that required participants to judge whether a randomly generated pattern was symmetrical along a vertical centerline. In the encoding component, each target location appeared as a red square on the grid for 650 ms before disappearing. The distractor symmetry component showed randomly generated black-and-white patterns, manipulated to be either symmetric or asymmetric along the vertical centerline. The order of sequences, target locations, and symmetry patterns were all randomized.

Participants had limited time to make symmetry judgments, with the maximum decision time determined during a pre-measurement phase. In this phase, participants completed 15 self-paced symmetry judgments, and the time limit was calculated as the mean + 2.5 SD of the last 12 trials. At the end of each sequence, participants were prompted to recreate the sequence of targets on an empty 4×4 grid.

Bandit Task

In the spatially correlated bandit task[21], participants completed 14 blocks of 25 trials, where they were instructed to maximize the number of points, which would be later converted to real money. Participants earned points by clicking tiles on a 11×11 square grid, and were explicitly informed they could also reclick previously revealed tiles. Each block consisted of 25 free-choice trials to sample arms (i.e., tiles on the grid) and collect their rewards, where the grid started off empty except for one randomly revealed tile. Since the number of trials (25) was significantly less than the total number of options (121), efficient exploration required generalization from limited experience.

To provide traction for generalization, participants were informed that nearby tiles tended to yield similar rewards, but the exact structure was unknown. Reward information was displayed using both tile colors and numbers, where darker colors indicated higher reward. In the LL condition, the most recent payout from each visited tile remained visible until the end of the block. However, in the HL condition, reward information disappeared after 400 ms, requiring participants to rely memory.

We randomized the order of blocks for each participant. To minimize the transfer of information between blocks, we also randomized the range of rewards in each grid using the following procedure. First, we pregenerated a set of 50 environments by sampling functions from a standard Gaussian Process prior with the mean function set to zero and using a radial basis function (RBF) kernel (Eq. 3) with the lengthscale parameter $\lambda = 2$. Then, we sampled without replacement 14 environments for each participant. Finally, each environment was normalized into a $[0,1]$ range and then scaled by a random uniform variable $\sim U(65,85)$.

Analyses

All significance tests were non-directional, with the significance level set to 0.05.

Behavioral analyses

All analyses including the reward variable used normalized rewards. Each observed reward value was normalized via feature scaling relative to the minimum and maximum expected reward of the corresponding bandit environment. Thus, a normalized reward of 0 signifies the minimum (expected) reward of a bandit, and 1 the maximum. In practice, due to the environment stochasticity, observed rewards could be fall outside the min-max range. Values above the max were clamped to 1 and values below the min were clamped to 0. For models reported in Results, normalized rewards were mean-centered within participants.

The working memory (WM) score was computed for each participant using partial-credit scoring[44]. This can be understood as a two-step calculation. First, we calculated the proportion of correct responses (i.e., correct spatial and sequence position) for each of the 2×4 recall blocks, and then calculated the mean of these proportions. We used z-scored WM scores in statistical analyses.

Computational modeling

For each model type, we fitted a vector of parameter values to each participant's data from each condition. The parameters were fitted using leave-one-block-out cross-validation (CV). For each of the 7 CV folds, we estimated maximum-likelihood parameter values using the differential evolution algorithm. The population size was set to 10 times the number of parameters being optimized. The mutation factor (F) was set to 0.8. We used random selection with the binomial crossover (with $p = .5$). The algorithm stopped upon reaching the convergence criterion of absolute difference of 10^{-3}.

To compare the models, we used the CV loss (negative log likelihood) on the held-out block, aggregated across all 7 CV folds. The final parameter estimates were obtained by taking the unweighted mean across the estimates from all CV folds.

Cross-validation loss was defined as the negative log-likelihood of the parameter vector θ_M (M indexing the model type):

$$\text{CV Loss} = -\log \mathscr{L}(\theta_M) = -\sum_{t=1}^{25} \log p_{\theta_M}(\mathbf{x}_{j,t} \mid \mathscr{D}_t),$$

where $p_{\theta_M(\mathbf{x}_t|\mathscr{D}_t)}$ is the probability of observation $\mathbf{x}_t$ given by model θ_M conditional on data up to trial t ($\mathscr{D}_t$).

The parameters were optimized within bounded regions: $\lambda, \beta, \tau \in [e^{-10}, e^5]$; $w_\mathrm{r} \in [e^{2.5}, e^5]$; $w_\mathrm{s} \in [e^{1.25}, e^{3.125}]$; $\alpha \in [e^{1.875}, e^5]$. In practice, we defined search bounds in the log space (e.g., $[-10,5]$ instead of $[e^{-10}, e^5]$) and exponentiated the sampled proposals for simulations. Parameter bounds were determined to ensure robust model recoverability (see Fig. S4).

Software

Statistical analyses were performed in `R v4.5.0`. Generalized linear mixed-effects models were fit using `glmmTMB v1.1.12`; general linear mixed-effects models were fit using `lme4 v1.1-37`. Hypothesis t-tests on model coefficients from `lme4` models were performed using `lmerTest v3.1-3`; the z-tests on model coefficients from `glmmTMB` models were included in the package; all other tests were performed using `R`'s base package.

Computational modeling tasks (simulations and parameter fitting) were performed in `julia v1.10.3`. Gaussian Process regression relied on `AbstractGPs v0.5.24` and `KernelFunctions v0.10.65`. Differential evolution was implemented in `Evolutionary v0.11.1`. Data visualizations were performed using `GLMakie v0.13.5`.

References

1. Shiffrin, R. M. Capacity limitations in information processing, attention, and memory. *Handb. learning cognitive processes* **4**, 177–236 (1976).
2. Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. review* **63**, 81 (1956).
3. Ma, W. J., Husain, M. & Bays, P. M. Changing concepts of working memory. **17**, 347–356, DOI: 10.1038/nn.3655.
4. Hasher, L. & Zacks, R. T. Working memory, comprehension, and aging: A review and a new view. *Psychol. learning motivation* **22**, 193–225 (1988).
5. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. neuroscience* **21**, 1609–1617 (2018).
6. Myers, N. E., Stokes, M. G. & Nobre, A. C. Prioritizing information during working memory: beyond sustained internal attention. *Trends cognitive sciences* **21**, 449–461 (2017).
7. Sakaki, M., Fryer, K. & Mather, M. Emotion strengthens high-priority memory traces but weakens low-priority memory traces. *Psychol. Sci.* **25**, 387–395 (2014).
8. Ying, Z., Callaway, F., Kiyonaga, A. & Mattar, M. G. Resource-rational encoding of reward information in planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46 (2024).
9. Ebbinghaus, H. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* (Duncker & Humblot, Leipzig, 1885).
10. Baddeley, A. D. & Hitch, G. The recency effect: Implicit learning with explicit retrieval? **21**, 146–155, DOI: 10.3758/BF03202726.
11. Lemaire, B. & Portrat, S. A computational model of working memory integrating time-based decay and interference. *Front. psychology* **9**, 416 (2018).
12. Zhou, H., Bamler, R., Wu, C. M. & Tejero-Cantero, A. Predictive, scalable and interpretable knowledge tracing on structured domains. In *The Twelfth International Conference on Learning Representations*, DOI: 10.48550/arXiv.2403.13179 (2024).
13. Rouhani, N. & Niv, Y. Signed and unsigned reward prediction errors dynamically enhance learning and memory. *Elife* **10**, e61077 (2021).
14. Koch, C., Zika, O., Bruckner, R. & Schuck, N. W. Influence of surprise on reinforcement learning in younger and older adults. *PLOS Comput. Biol.* **20**, e1012331 (2024).
15. Gershman, S. J. & Daw, N. D. Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annu. Rev Psychol* **68**, 101–128 (2017).
16. Botvinick, M. *et al.* Reinforcement learning, fast and slow. *Trends cognitive sciences* **23**, 408–422 (2019).
17. Wu, C. M., Meder, B. & Schulz, E. Unifying principles of generalization: past, present, and future. *Annu. Rev. Psychol.* **76**, 275–302, DOI: 10.1146/annurev-psych-021524-110810 (2025).
18. Wimmer, G. E., Daw, N. D. & Shohamy, D. Generalization of value in reinforcement learning by humans. *Eur. J. Neurosci.* **35**, 1092–1104 (2012).
19. Biderman, N. & Shohamy, D. Memory and decision making interact to shape the value of unchosen options. *Nat. communications* **12**, 4648 (2021).
20. Rasmussen, C. E. & Williams, C. K. I. *Gaussian processes for machine learning*. Adaptive computation and machine learning (MIT Press, Cambridge, Mass, 2006). OCLC: ocm61285753.
21. Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D. & Meder, B. Generalization guides human exploration in vast decision spaces. **2**, 915–924, DOI: 10.1038/s41562-018-0467-4.
22. Giron, A. P. *et al.* Developmental changes in learning resemble stochastic optimization, DOI: 10.31234/osf.io/9f4k3.
23. Wu, C. M., Schulz, E., Garvert, M. M., Meder, B. & Schuck, N. W. Similarities and differences in spatial and non-spatial cognitive maps. *PLOS Comput. Biol.* **16**, 1–28, DOI: 10.1371/journal.pcbi.1008149 (2020).
24. Wu, C. M., Schulz, E. & Gershman, S. J. Inference and search on graph-structured spaces. *Comput. Brain & Behav.* **4**, 125–147 (2021).
25. Bhui, R., Lai, L. & Gershman, S. J. Resource-rational decision making. *Curr. Opin. Behav. Sci.* **41**, 15–21 (2021).
26. Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. **35**, 1024–1035, DOI: 10.1111/j.1460-9568.2011.07980.x (2012).
27. Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A. & Frank, M. J. Working memory contributions to reinforcement learning impairments in schizophrenia. **34**, 13747–13756, DOI: 10.1523/JNEUROSCI.0989-14.2014 (2014).
28. Collins, A. G., Albrecht, M. A., Waltz, J. A., Gold, J. M. & Frank, M. J. Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia. *Biol. psychiatry* **82**, 431–439 (2017).
29. Montaser-Kouhsari, L., Nicholas, J., Gerraty, R. T. & Shohamy, D. Differentiating reinforcement learning and

episodic memory in value-based decisions in parkinson's disease. *J. Neurosci.* **45** (2025).

30. Baddeley, A. The episodic buffer: a new component of working memory? *Trends cognitive sciences* **4**, 417–423 (2000).

31. Greene, N. R. & Naveh-Benjamin, M. Adult age-related changes in the specificity of episodic memory representations: A review and theoretical framework. *Psychol. Aging* **38**, 67 (2023).

32. Cogliati Dezza, I., Cleeremans, A. & Alexander, W. Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *J. Exp. Psychol. Gen.* **148**, 977–993 (2019).

33. Wu, C. M., Schulz, E., Pleskac, T. J. & Speekenbrink, M. Time pressure changes how people explore and respond to uncertainty. *Sci Rep* **12**, 4122 (2022).

34. Brown, V. M., Hallquist, M. N., Frank, M. J. & Dombrovski, A. Y. Humans adaptively resolve the explore-exploit dilemma under cognitive constraints: Evidence from a multi-armed bandit task. *Cognition* **229**, 105233 (2022).

35. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *J. experimental psychology: Gen.* **143**, 2074 (2014).

36. Gershman, S. J. Uncertainty and exploration. **6**, 277–286, DOI: 10.1037/dec0000101. Place: US Publisher: Educational Publishing Foundation.

37. Radulescu, A., Niv, Y. & Ballard, I. Holistic reinforcement learning: the role of structure and attention. *Trends cognitive sciences* **23**, 278–292 (2019).

38. Le, Q. V., Smola, A. J. & Canu, S. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, 489–496 (2005).

39. Witt, A., Toyokawa, W., Lala, K. N., Gaissmaier, W. & Wu, C. M. Humans flexibly integrate social information despite interindividual differences in reward. *Proc. Natl. Acad. Sci.* **121**, e2404928121 (2024).

40. Nagy, D. G., Orban, G. & Wu, C. M. Adaptive compression as a unifying framework for episodic and semantic memory. *Nat. Rev. Psychol.* DOI: 10.1038/s44159-025-00458-6 (2025).

41. Foster, J. L. *et al.* Shortened complex span tasks can reliably measure working memory capacity. *Mem. & cognition* **43**, 226–236 (2015).

42. Oswald, F. L., McAbee, S. T., Redick, T. S. & Hambrick, D. Z. The development of a short domain-general measure of working memory capacity. *Behav. research methods* **47**, 1343–1355 (2015).

43. Ten, A. *et al.* Forgetful generalization, DOI: 10.17605/OSF.IO/FDUXB (2025).

44. Conway, A. R. *et al.* Working memory span tasks: A methodological review and user's guide. *Psychon. bulletin & review* **12**, 769–786 (2005).

45. Bonawitz, E., Denison, S., Gopnik, A. & Griffiths, T. L. Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. **74**, 35–65, DOI: 10.1016/j.cogpsych.2014.06.003.

46. Botvinick, M. *et al.* Reinforcement learning, fast and slow. **23**, 408–422, DOI: 10.1016/j.tics.2019.02.006.

47. Wu, C. M., Schulz, E. & Gershman, S. J. Inference and search on graph-structured spaces. **4**, 125–147, DOI: 10.1007/s42113-020-00091-x.

48. Knowlton, B. J. & Castel, A. D. Memory and reward-based learning: A value-directed remembering perspective. **73**, 25–52, DOI: 10.1146/annurev-psych-032921-050951.

49. Witt, A., Toyokawa, W., Lala, K. N., Gaissmaier, W. & Wu, C. M. Humans flexibly integrate social information despite interindividual differences in reward. *Proc. Natl. Acad. Sci.* **121**, e2404928121, DOI: 10.1073/pnas.2404928121 (2024).

50. Altmann, E. M. & Schunn, C. D. Decay versus interference: A new look at an old interaction. *Psychol. Sci.* **23**, 1435–1437 (2012).

51. Modirshanechi, A., Brea, J. & Gerstner, W. A taxonomy of surprise definitions. *J. mathematical psychology* **110**, 102712 (2022).

52. Price, K. V., Storn, R. M. & Lampinen, J. A. *Differential evolution: a practical approach to global optimization* (Springer, 2005).

53. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. *Neuroimage* **84**, 971–985 (2014).

54. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife* **8**, e49547 (2019).

55. Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E. & Camos, V. Time and cognitive load in working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **33**, 570 (2007).

56. Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural computation* **7**, 108–116 (1995).

57. Tikhonov, A. N. Solutions of ill posed problems. (1977).

58. Findling, C. & Wyart, V. Computation noise in human learning and decision-making: origin, impact, function. *Curr. Opin. Behav. Sci.* **38**, 124–132 (2021).

59. Callaway, F. *et al.* Rational use of cognitive resources in human planning. *Nat. human behaviour* **6**, 1112–1125 (2022).

60. Gershman, S. J. The rational analysis of memory. *Oxf. handbook human memory.* (2021).

61. Popov, V., Marevic, I., Rummel, J. & Reder, L. M. Forgetting is a feature, not a bug: Intentionally forgetting some things helps us remember others by freeing up working memory resources. *Psychol. Sci.* **30**, 1303–1317 (2019).

62. Craik, F. I., Govoni, R., Naveh-Benjamin, M. & Anderson, N. D. The effects of divided attention on encoding and retrieval processes in human memory. *J. Exp. Psychol. Gen.* **125**, 159 (1996).

63. Unsworth, N. & Engle, R. W. The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychol. review* **114**, 104 (2007).

64. Beukers, A. O., Buschman, T. J., Cohen, J. D. & Norman, K. A. Is activity silent working memory simply episodic memory? *Trends cognitive sciences* **25**, 284–293 (2021).

65. Baddeley, A. D., Allen, R. J. & Hitch, G. J. Binding in visual working memory: The role of the episodic buffer. *Explor. Work. Mem.* 312–331 (2017).

66. Cabeza, R., Dolcos, F., Graham, R. & Nyberg, L. Similarities and differences in the neural correlates of episodic memory retrieval and working memory. *Neuroimage* **16**, 317–330 (2002).

67. Lugtmeijer, S., de Haan, E. H. & Kessels, R. P. A comparison of visual working memory and episodic memory performance in younger and older adults. *Aging, Neuropsychol. Cogn.* **26**, 387–406 (2019).

68. Valenti, L. & Galera, C. Dynamic visual noise has the same effect on visual memory and visual imagery tasks. *Psychol. & Neurosci.* **13**, 114 (2020).

69. Yin, H., Wang, Y., Zhang, X. & Li, P. Feedback delay impaired reinforcement learning: Principal components analysis of reward positivity. *Neurosci. letters* **685**, 179–184 (2018).

70. Rubino, V., Hamidi, M., Dayan, P. & Wu, C. M. Compositionality under time pressure. In Goldwater, M., Anggoro, F., Hayes, B. & Ong, D. (eds.) *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, DOI: 10.31234/osf.io/z2648 (Cognitive Science Society, Sydney, Australia, 2023).

71. Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. *Applied multiple regression/correlation analysis for the behavioral sciences* (Routledge, 2013).

72. Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J. & Hilbig, B. E. lab. js: A free, open, online study builder. *Behav. Res. Methods* **54**, 556–573 (2022).

73. Lange, K., Kühn, S. & Filevich, E. " just another tool for online studies"(jatos): An easy solution for setup and management of web servers supporting online studies. *PloS one* **10**, e0130834 (2015).

74. de Holanda Coelho, G. L., Hanel, P. H. P. & Wolf, L. J. The very efficient assessment of need for cognition: Developing a six-item version. *Assessment* **27** (2018).

75. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *elife* **8**, e49547 (2019).

Acknowledgements

CMW was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645, and funded by the DFG under Germany's Excellence Strategy – EXC 2117 – 422037984. The authors acknowledge support by the state of Baden-Württemberg through bwHPC (Baden-Württemberg High Performance Computing) and thank the Self-Regulation Hub at Hector Research Institute for helpful feedback and discussion.

Author contributions statement

Conceptualization: C.M.W., S.B., A.T., K.M., M.S.; Methodology: C.M.W., S.B., A.T.; Software: A.T., S.B., C.M.W., A.C.; Formal analysis: A.T., C.M.W.; Writing – original draft: A.T., C.M.W.; Writing – review & editing: A.T., C.M.W., K.M., M.S.; Supervision: C.M.W., K.M., M.S.

Additional information

The code for computational modeling, statistical analyses, data visualization and processing is available at `https://anonymous.4open.science/r/gp-E172`.

The authors declare no competing interests.

Supplementary Information for
In search of lost memories: Modeling exploration with forgetful generalization

Ten, A., Sakaki, M., Breit, S., A. Chandrasekaran, Murayama, K., & Wu, C.M.

Effect of working memory on performance

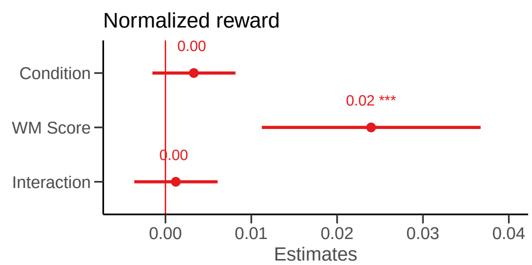

Figure S1. Model coefficients from `reward ∼ cond * WMScore + (cond|id)`; *** indicates $p < .001$.

Modulation of reward sensitivity by memory load

The preregistered Poisson model indicated a non-significant modulation of reward sensitivity by condition. However, the raw data in Fig. S2a,b suggested a slight downward trend in the effect of condition on search distance as rewards increased. Compared to the linear model, the Poisson model had much larger confidence intervals for estimated means at lower levels of the reward. This is not surprising, given that participants were instructed to maximize earnings and thus avoided low-rewarding choices. Due to the tighter confidence intervals in the linear model, the interaction between reward and condition appears significant (Fig. S2c). However, model comparisons between the Poisson and linear models proved inconclusive (see Fig. S2 caption).

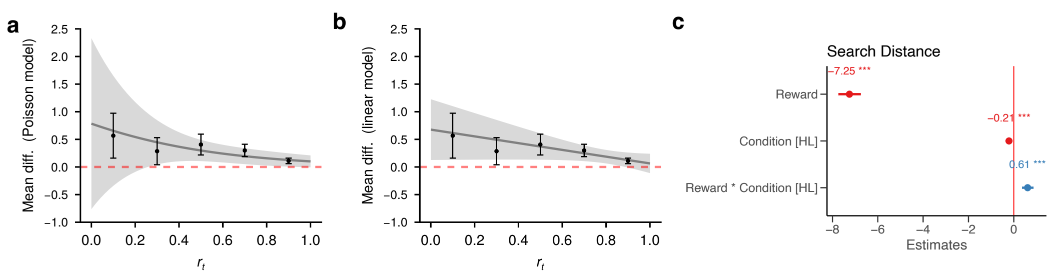

Figure S2. **a**) Inset from Fig. 3e of the main text. The plot is showing for each reward bin, the mean differences (±95% CI) in search distance between LL and HL conditions (positive values indicate higher distance in LL). Raw data is represented by dots and whiskers, and the mean±95% CI predicted by the Poisson mixed-effects model `dist ∼ reward * cond + (reward * cond|id)` is represented as gray line and band. **b**) Shows the same raw data as **a**, but the predictions (gray line and band) are given by a linear model (with Gaussian likelihood and an identity link). The linear model performed slightly better in terms of 30-fold cross-validated MSE (MSE_{Linear} $[95\%\ CI] = 4.40\ [4.29,\ 4.52]$, compared to MSE_{Poisson} $[95\%\ CI] = 5.23\ [5.00,\ 5.45]$), but the information criteria strongly favored the Poisson model ($\Delta AIC_{\text{Poisson}-\text{Linear}} = -62203.23$, $\Delta BIC_{\text{Poisson}-\text{Linear}} = -62212.37$). We note, however, that the information criteria may not be appropriate for comparing these models, since their likelihood functions are qualitatively different. Based on these investigations, we cannot draw strong conclusions about the modulatory effect of load condition on reward sensitivity. **c**) Model coefficients the linear model; *** indicates $p < .001$.

Working memory and condition-modulated reward sensitivity

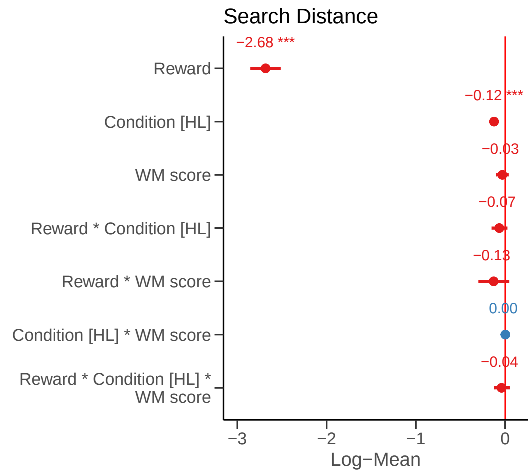

Figure S3. Model coefficients from `dist ~ reward * cond * WMScore + (reward * cond|id)`; *** indicates $p < .001$.

Computational modeling

Model recovery

Using each model type $M \in \{\text{GP}, \text{GPR}, \text{GPS}, \text{GPS}_\text{a}, \text{GPRS}, \text{GPRS}_\text{a}\}$, we simulated 96 participants and their synthetic data (we determined the number of simulated participants based on the maximum number of CPU jobs available on the computer cluster). Each synthetic participant was generated by sampling a parameter vector with $n(M)$ elements (e.g., $n(\text{GP}) = 3$ and $n(\text{GPRS}_\text{a}) = 6$). The baseline parameter values were sampled from a multivariate Gaussian distribution fitted to the sample of parameters obtained in Witt et al.[39] (solo condition), which used a setup similar to our control (low load) condition, except for a shorter horizon (15 trials). When present in the model, the "forgetfulness" parameters (w_r, w_s, and $alpha$) were sampled from a log-uniform distribution between e^{-5} and e^5. In practice, we sampled values uniformly between -5 and 5 and exponentiated them when simulating the data. After fitting the models, we constructed the confusion (Fig. S4a) and inversion matrices (Fig. S4b) to assess model recovery[75].

As demonstrated by Fig. S4a, models defined naively over broad parameter regions cannot be reliably recovered. The problem is the nested structure of the models, e.g., GP is a special case of GPRS with $w_\text{r} = w_\text{s} = 0$. To address this, we performed a grid search in the forgetfulness parameters that would enable us to recover the generating model at a satisfactory rate. The idea was that if our model comparison procedure failed to recover, for example, the generating GPR model by confusing it with the simpler GP model, this would be due to the generating w_r parameter being effectively indistinguishable from 0. If the model has internal validity, there should be a region of w_r at which the generated data becomes more distinct from what the baseline GP model generates. The same logic applies to the GPS model and the w_s parameter.

Our search for good recoverability regions proceeded as follows. First, we generated 96 synthetic baseline participants as before. These are equivalent to participants with $w_\text{r} = w_\text{s} = 0$. Then, separately for models GPR and GPS, we manipulated either w_r or w_s of each participant to take on one of 16 log-linearly spaced values on the interval $[e^{-5}, e^5]$. Thus, we created 16×2 mutated clones of each synthetic participant, each of which had either w_r or w_s altered from 0 to a non-zero value $\in \{e^{-5+0.625 \cdot i} \mid i = 0, \ldots, 15\}$. We then simulated 7 blocks of 25 choices with each participant, resulting in $96 + 96 \times 16 \times 2$ datasets. Finally, separately for each model type (GPR and GPS), we fitted both the baseline and a corresponding forgetful model to all datasets to assess recovery. Specifically, we fitted the baseline model to each of the baseline-generated datasets and each of the forgetful-model-generated datasets. We then fitted the forgetful models to baseline datasets and to their respective forgetful-model-generated datasets. When fitting the forgetful models, we manipulated the lower bound of the optimization interval for the corresponding forgetful parameter. Thus, each of the 96 baseline datasets was fitted by GPR model 16 times, once inside the interval $[e^{-5}, e^5]$ for the w_r parameter, once inside $[e^{-4.375}, e^5]$, and so on until $[e^{4.375}, e^5]$. The procedure was replicated for the GPS model and the w_s parameter. When fitting forgetful models to the forgetful-model-generated data, the lower bound was set to the same level as the true generating parameter.

The results are presented in Fig. S4b. Recoverability of the baseline and the forgetful GPR models increased steadily as a function of w_r, going beyond 90% for both at $e^{2.5}$. Thus, when the data from the forgetful model was generated by either the

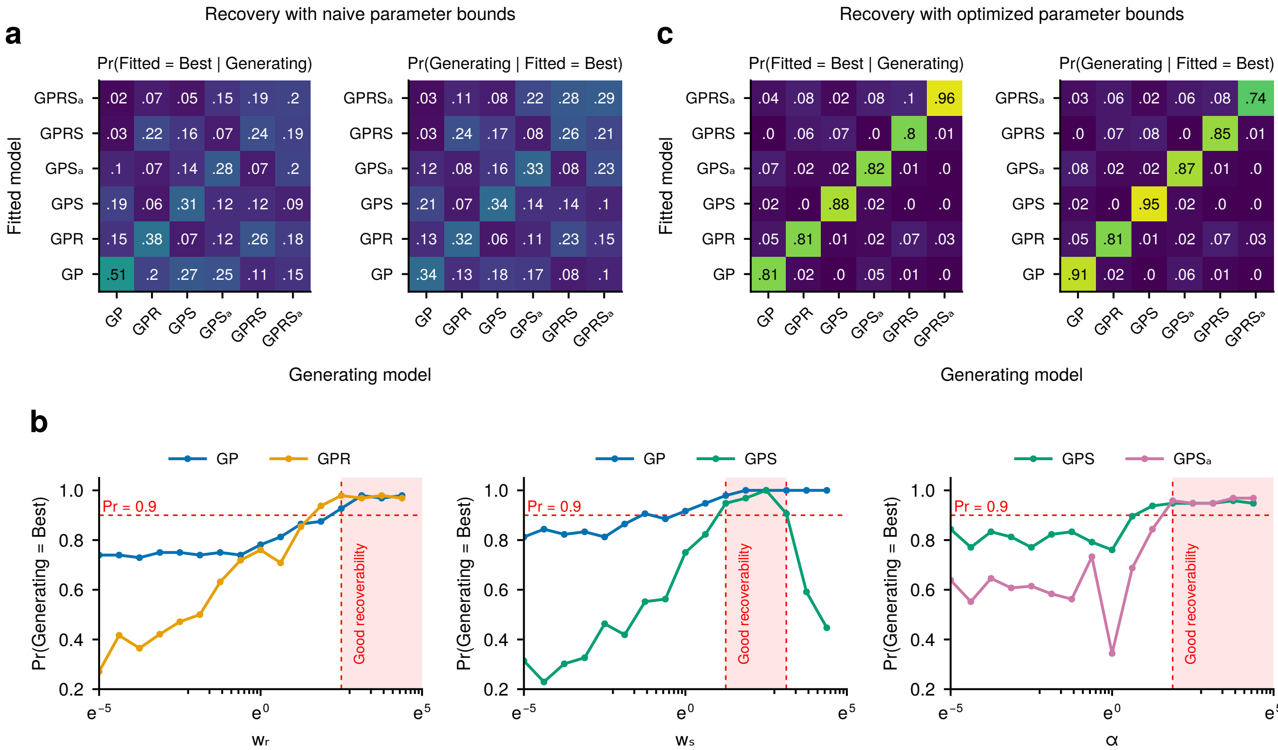

Figure S4. **Model recovery. a**) Confusion (left) and inversion (right) matrices. The confusion matrix shows the the estimated probability that a fitted model (listed in rows) fits better than other models, given the data generated from a specific generating model (organized in columns). The inversion matrix gives us probabilities of a generating model, given a certain best fitting model. Using naive parameter bounds on the distributions of generating parameters and the search bounds of differential evolution, generating models cannot be reliably recovered or identified. **b**) Illustrates our bound restriction procedure to ensure model parameters result in distinct behavioral patterns which can reliably identify the generating models. The lines show the estimated probability that generating model (e.g., GP on the left most plot) is better than the competing model (GPR or the left most plot). As we increase the value of the lower bound of the w_r parameter in the GPR model, the probability of correctly selecting each of the models increases and reaches a satisfactory level of .9 at $w_r = e^{2.5}$. (red region). The same procedure was applied to set the lower and upper bounds on the w_s parameter (middle plot) that allow us to reliably distinguish between GP and GPS models. Finally, the procedure was repeated again to find the lower bound on the asymmetry α parameter that allows to differentiate between the GPS and GPS' models (both with informed w_s bounds). **c**) confusion (left) and inversion (right) matrices for forgetful models with recover-informed parameter bounds. Constraining the generating and fitted parameters to specific regions allowed us to define models that were highly distinct from each other.

baseline model or the forgetful model with $w_r > e^{2.5}$, we were able to recover the generating model at least 90% of the time. Thus, we set the fixed the lower bound of the optimization search for the w_r parameter at $e^{2.5}$. We left the upper bound at 5 for this parameter, as recoverability was still above 90% at that level. Similarly, recoverability of the baseline and the forgetful GPS models increased as a function of w_s, however at high levels of w_s, it worsened again. Using the same approach, we bounded the w_s parameter search in $[e^{1.25}, e^{3.125}]$.

The GPS is a special case of GPS_a, since any GPS model is equivalent to GPS_a with $\alpha = 1$. The issue is similar to that raised by the nestedness of the GP model within GPR and GPS. To address this issue, we performed another set of recoverability optimization, this time focusing on the asymmetry parameter α. Specifically, we generated 96 synthetic forgetful symmetric-surprise-sensitive participants using the $[e^{1.25}, e^{3.125}]$ interval for w_s established earlier. Next, we mutated each of these participants by setting the asymmetry parameter α to a value $\in \{e^{-5+0.625 \cdot i} \mid i = 0, \ldots, 15\}$. Next, we fitted the GPS and GPS$'$ models to the data generated by the forgetful participants and their mutated counterparts. As before, we set the lower bound for parameter fitting of the α parameter according to the generating value. The results are shown in Fig. S4b. We could only differentiate positively-asymmetric GPRS_a models from the GPS that were satisfactorily distinct from the baseline GP model. As expected, models with the asymmetry parameter set at 0 were the most difficult to recover when compared to simpler GPS models. The recovery rate for both model types crossed our criterion of 90% at $\alpha = e^{3.125}$. Thus, the lower bound the α parameter in asymmetric models was set to $e^{3.125}$.

When we generated data using these optimized parameter bounds, the recovery improved dramatically (Fig. S4c). Restricting parameter bounds allows us to be more confident in our model comparison results, because, assuming that the data was generated by one of the 6 models, we can be 74-95% confident that the generating model is indeed the best fitting model.

Parameter recovery

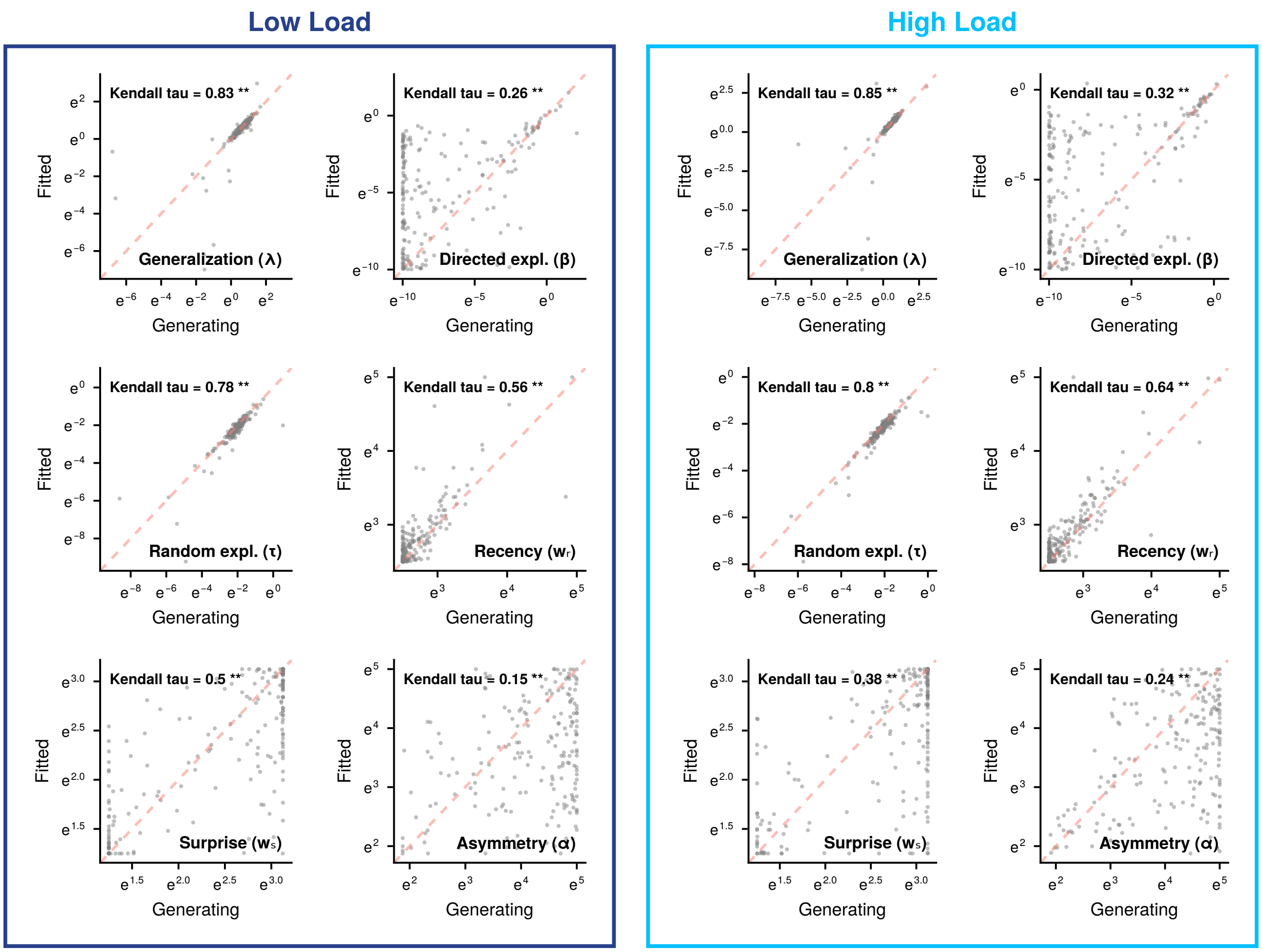

Figure S5. Parameter recovery. Each subplot illustrates the joint distribution of the generating and fitted parameter values from the GPRS$_a$ model. Note the log scale for both. The red dashed diagonal lines signify identity.

Parameter recovery is summarized in Fig. S5. To assess recovery, we first simulated choices across 7 blocks of 25 trials using the GPRS$_a$ models fitted to each participant in each condition. We then used the same parameter fitting procedure on the simulated data as we used on the human data to get the recovered values. The recovery patterns were similar across conditions. all fitted parameters correlated significantly ($p < .01$) with the generating parameters, indicating that parameter values have meaningful effects on observable choices. Recovery of generalization (λ) and random-exploration (τ) parameters was excellent (Kendall's $\tau \in [.78\ .85]$); recovery of recency (w_r) and surprise (w_s) prioritization parameters was good (Kendall's $\tau \in [.38,\ .64]$); recovery of directed exploration (β) and surprise-asymmetry (α) parameters was modest, but statistically significant (Kendall's $\tau \in [.15,\ .32]$).

For reference, we also include histograms of parameter estimates for each model type in Fig. S8.

Computational model parameters predicted by memory load and WM score

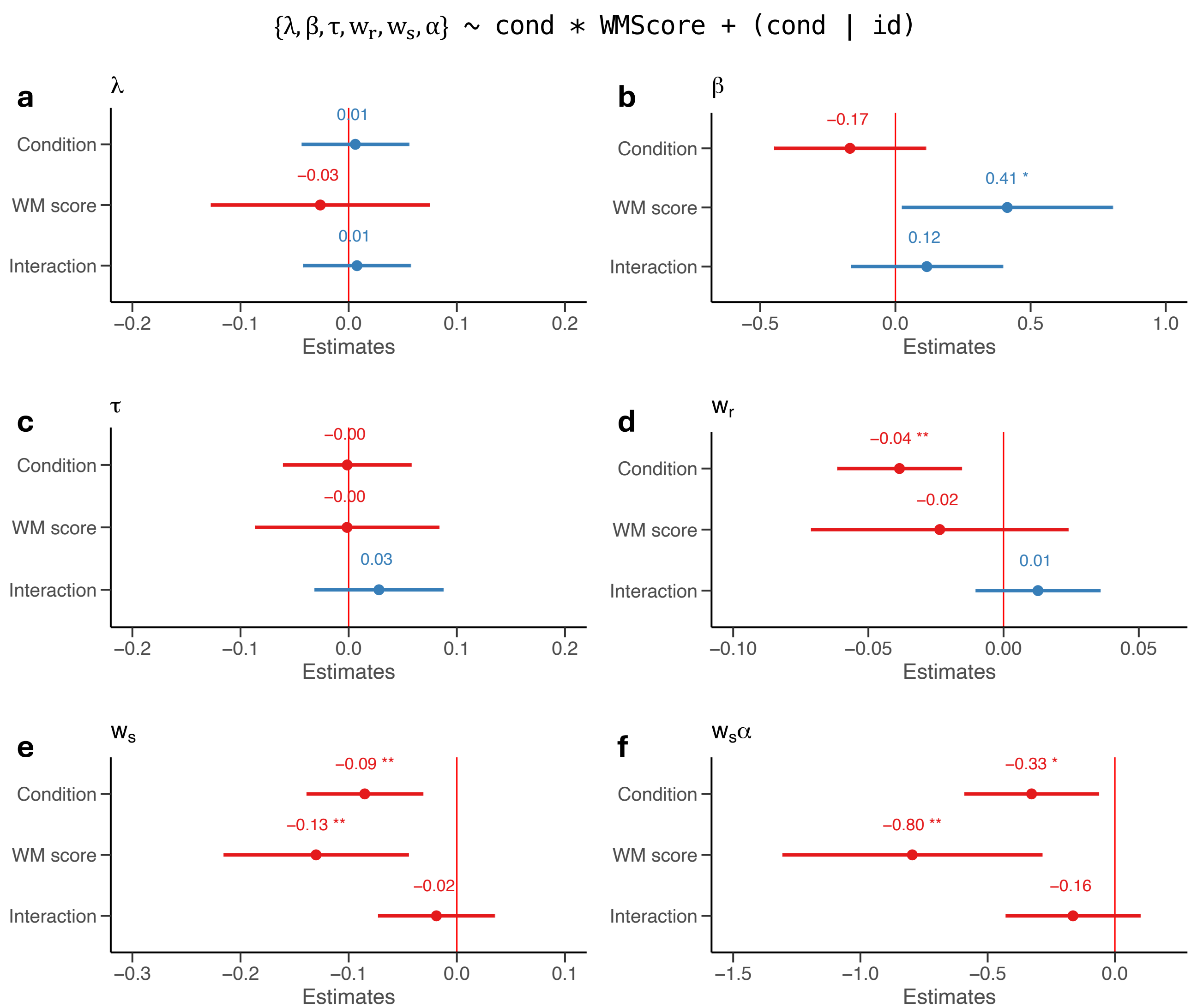

Figure S6. Coefficients from linear mixed-effects regressions of GPRS_a model parameters on condition and WM score. *** indicates $p < .001$; ** indicates $p < .01$; * indicates $p < .05$.

Winning model (GPRS_a) parameters and human behavior

For these analyses, we turned to three behavioral indices: mean normalized reward, mean search distance, and mean reward sensitivity. Mean normalized reward was calculated as the mean normalized reward in each condition and each participant. Mean search distance was defined as the Manhattan distance between consecutive choices averaged for each condition in each participant. Mean reward sensitivity was calculated as negative Kendall's τ statistic – quantifying the association between reward at time t and search distance at $t+1$ – again, separately for each condition in each participant. We then regressed each of these behavioral indices on the interaction between fitted parameters and condition, using a linear mixed-effects model. The general form was `beh ~ param * cond + (param * cond | id)`, where `beh` was either mean normalized reward or mean reward sensitivity, and `param` was one of the $\{\lambda, \beta, \tau, w_\text{r}, w_\text{s}, \alpha \cdot w_\text{s}\}$. We analyzed the 12 resulting models to detect significant associations between the fitted model parameters and behavior. The results are summarized in Fig. S7.

Mean normalized reward, as an index of individual performance, was predicted by every parameter of the GPRS_a model

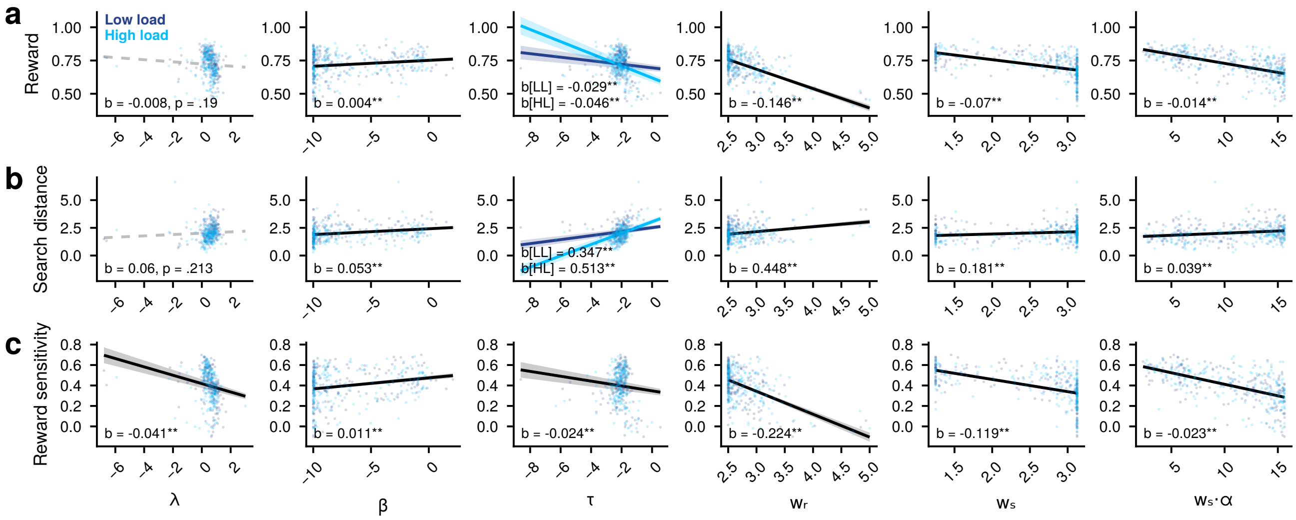

Figure S7. Model parameters predict behavior. Rows correspond to behavioral indices (**a**: reward, **b**: search distance, **c**: reward sensitivity), and columns correspond to model parameters (from left to right, generalization (λ), directed exploration (β), random exploration (τ, recency prioritization (w_r, negative surprise prioritization (w_s, and positive surprise prioritization ($w_s \cdot \alpha$). Each panel shows a joint distribution of fitted values of model parameters and behavioral indices (conditions superimposed). The lines illustrate the effects. Where both the main effect of parameter and and interaction with condition are significant ($p < .01$), we plot predictions separately for each condition (in light and dark blue). Where the overall effect of parameter on behavior was significant ($p < .01$) in the absence of interaction ($p \geq .01$), we show a single black line. Nonsignificant overall effects are shown as gray dashed line. Effect estimates are included in each subplot; double asterisk (**) indicates $p < .01$. either for overall effect, or for both the overall and the interaction effects.

except the generalization parameter λ (Fig. S7a). Performance was positively correlated with directed exploration (β) and negatively correlated with random exploration (τ), as well as memory prioritization weights (w_r, w_s, and $w_s \cdot \alpha$ (all $p < .01$; see Fig. S7 for coefficients). Moreover, the effect of random exploration was significantly stronger in the High load condition. Overall, these results suggest that undirected exploration and memory prioritization were costly for performance.

Search distance was also associated with every parameter of the model, except the generalization parameter λ (Fig. S7b). All correlations were positive, indicating that exploration (both directed and random), as well as memory prioritization corresponded to less local search.

Reward sensitivity was predicted by all parameters predicted positively, with the exception of directed exploration (β). Specifically, increased reward sensitivity was associated with less generalization, less random exploration (and more directed exploration), and less memory prioritization (Fig. S7c).

Distributions of fitted model parameters from all models

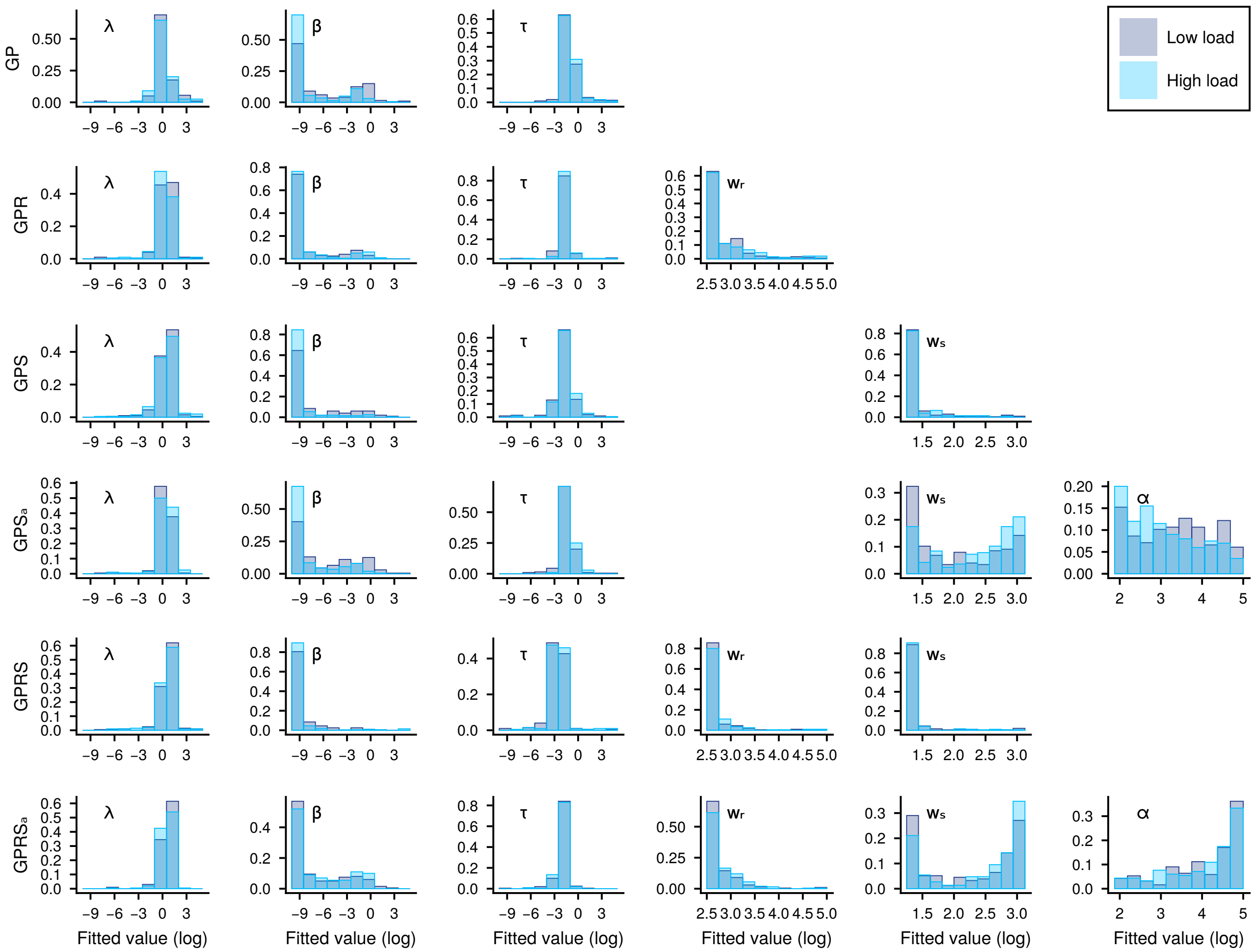

Figure S8. Parameter estimates. Distributions of fitted parameters for each model type. Model types occupy rows and parameters occupy columns. Note that the *x* axis corresponds to the logarithms of the fitted values.