npj | science of learning

Article

Published in partnership with The University of Queensland

https://doi.org/10.1038/s41539-025-00363-w

The distinct functions of working memory and intelligence in model-based and model-free reinforcement learning

Chengyan Yang[1,2], Tongran Liu[1,2] ✉, Mengxin Wen[1,2] & Xun Liu[1,2]

Human and animal behaviors are influenced by goal-directed planning or automatic habitual choices. Reinforcement learning (RL) models propose two distinct learning strategies: a model-based strategy, which is more flexible but computationally demanding, and a model-free strategy is less flexible yet computationally efficient. In the current RL tasks, we investigated how individuals adjusted these strategies under varying working memory (WM) loads and further explored how learning strategies and mental abilities (WM capacity and intelligence) affected learning performance. The results indicated that participants were more inclined to employ the model-based strategy under low WM load, while shifting towards the model-free strategy under high WM load. Linear regression models suggested that the utilization of model-based strategy and intelligence positively predicted learning performance. Furthermore, the model-based learning strategy could mediate the influence of WM load on learning performance. These findings underscore the critical role of WM capacity in strategic selection during RL process.

Learning to make choices based on feedback and to maximize rewards is crucial for survival. Two distinct systems have been identified in the decision-making process: a controlled system and an automatic system. The controlled system relies on goal-directed planning, which entails greater computational demands, while the automatic system depends on habitual processing that requires fewer computational demands[1–4]. It remains less understood whether these different systems are variably influenced by complex environments or individual mental capabilities.

Reinforcement learning (RL) has been characterized as a form of dopamine-dependent plasticity that shapes neural pathways to facilitate value-based learning within the brain[5]. To effectively plan and guide subsequent actions in response to environmental demands, organisms learn from accumulated experience. They strive to establish connections between their actions and outcomes while constructing cognitive models through modification of neural pathways in their brains. Based on extensive studies, the concept of reinforcement has expanded beyond traditional laws of effect to encompass environmental and state-based information, closely aligning with theories of instrumental conditioning observed in animals. Rodent behaviors in dynamic environments characterized by stochastic reward-punishment contingencies can be qualitatively simulated using RL algorithms that incorporate environmental information[6]. This demonstrates that knowledge about environmental state is as critical as reward signals in shaping behavior. Moreover, both humans and animals have demonstrated temporal difference learning and model-like valuations[1,3]; additionally, the neural substrates and value estimation mechanisms that governing goal-directed versus habitual behaviors have also been identified[1,4].

Two distinct RL strategies have been proposed by Daw and his colleagues: the model-based strategy and the model-free strategy[2,7,8]. The model-based RL strategy is characterized by its flexibility and is typically driven by specific goals, taking current states into account. This approach involves exploring the current environmental states and attempting to establish internal models based on environmental information to guide subsequent actions; however, this comes at the cost of efficiency due to increased computational requirements. In contrast, the model-free strategy is more efficient and resource-conserving; it tends to be driven by habits and immediate rewards while primarily linking current feedback with particular actions or stimuli without considering underlying environmental structures. The model-free strategy relies predominantly on prior action-outcome experiences, enabling the automatic repetition of reward-seeking behaviors with minimal cognitive demands compared to the model-based strategy[9]. Both human and animal studies have focused on elucidating the distinct roles played by model-based and model-free strategies in decision-making processes and learning mechanisms[9–15]. For instance, both humans and monkeys are capable of acquiring latent-state type strategies during two-step tasks, utilizing inference regarding the current latent state to guide their future behavior[16,17], thereby exhibiting a hybrid strategy that integrates both

[1]State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China. [2]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China. ✉e-mail: liutr@psych.ac.cn

model-based and model-free strategies throughout the learning process[9,18,19].

Traditional two-stage tasks have been employed to explore and differentiate between model-based and model-free RL strategies, wherein participants made choices between options that led to probabilistic transitions into different states in the subsequent stage[2,4,12,20–23]. Model-based and model-free strategies predicted distinct behavioral patterns regarding how rewards obtained in the second stage influenced first-stage choices in later trials. Numerous variations of this paradigm have emerged, incorporating diverse reward probabilities, transition structures, numbers of options and stages, as well as modifications to the instructions[24–30]. A novel two-stage task has been further developed and employed[25,27–30], which encompasses a broader range of reward quantities to enhance distinguishability among options while providing participants with more information. Furthermore, the deterministic transitions reduce randomness, thereby encouraging participants to better establish transition models and employ a model-based strategy. The choosing action in the original second stage is replaced by a simple collection action that prompts participants to focus more intently on their first-stage choice. The current study will adopt this innovative paradigm to further investigate how the utilization of model-based and model-free strategies is adjusted under varying mental loads, as well as how these strategies are influenced by other mental abilities, such as intelligence.

Working memory (WM) serves as an essential indicator of mental abilities. It refers to an information-limited process that enables individuals to temporarily hold representations in mind for subsequent thought and action, thereby allowing agents to store and manipulate information[31]. WM has been extensively shown to interact with RL processes[32–35]. Notably, WM capacity has been found to correlate with model-based decision-making in adults and may underpin RL processes[36]. According to dual-system RL theory, model-based learning heavily relies on environmental information. To sustain model-based learning, agents must allocate substantial cognitive resources to memorize the environmental structure and continuously verify and adjust their understanding based on their WM throughout the entire learning process. Moreover, WM acts as a mediating factor; individuals with lower WM capacity are more susceptible to detrimental effects of stress that impair their reliance on model-based learning, while those with higher WM ability exhibit a greater utilization of model-based strategies[23,36]. Zuo et al.[37] investigated the influence of WM processes on decision-making alongside underlying cognitive computing mechanisms. They further established the Hybrid-WM model and demonstrated a positive association between individuals' WM ability and model-based learning. However, it remains unclear how WM interacts with RL strategies to further influence learning performance. Considering the characteristics of RL, WM loads can manifest in both external and internal aspects of RL tasks. The external WM load may be perceived as an additional task, while the internal WM load could signify a more intricate environmental structure within the RL task. The effect of internal WM load on strategies and performance in RL tasks has been infrequently examined in prior research.

Fluid intelligence represents another critical aspect of mental abilities; it refers to an individual's capacity for employing deliberate mental operations for reasoning and solving novel problems[38,39]. Numerous empirical and educational studies have demonstrated that fluid intelligence is correlated with human learning, especially in complex learning scenarios[40–42]. Nevertheless, only a limited number of studies have investigated the association between fluid intelligence and model-based versus model-free RL process. Schad et al.[43] found that individuals exhibiting higher weights for model-based learning displayed faster processing speeds. Developmental studies have reported significant correlations between fluid reasoning and model-based strategy use, further indicating that fluid reasoning mediated the relationship between age and model-based decision-making[23]. Nonetheless, it remains less known whether intelligence or its interactions with WM and RL strategies can affect learning performance.

The current study aims to address two primary objectives. The first objective is to investigate how different RL strategies are influenced by various mental abilities, such as WM and intelligence. The second objective is to explore the impact of RL strategies and mental abilities on learning performance. We manipulated WM loads by employing tasks that involved either two pairs or four pairs of stimuli. This approach allowed us to assess how WM and intelligence influenced model-based and model-free RL strategies under varying WM loads. Our experimental design increased the WM load by incorporating environmental content requiring memorization, thereby focusing on investigating the effects of internal WM load. In addition to widely utilized computational parameters for RL process, such as learning rate (which reflects the efficiency of value updating), eligibility trace (representing the influence of the second-stage outcomes on first-stage decisions in subsequent trials) and inverse temperature (indicating the exploitation-exploration trade-off based on different values of options), we primarily analyzed the mixing weight parameter to represent the model-free and model-based strategies. It was hypothesized that WM loads and intelligence would differentially affect learning strategies, and the influence of mental abilities on overall learning performance could be modulated by these learning strategies.

Results

We adopted a two-stage RL task adapted from Kool et al.'s paradigm[27] with low and high WM load conditions. Each trial consisted of two stages under both conditions. In the first stage, participants were required to make a choice that determined the subsequent state in the second stage. The transition structure between these two stages was initially unknown; thus, participants needed to construct the structure mentally through iterative trial-by-trial attempts. In the second stage, participants pressed a fixed key to collect rewards. The number of rewards followed an independent Gaussian random walk distribution. The primary objective for participants was to maximize their total rewards. Further details regarding the experimental paradigm can be found in the Method section. Participants' performance was assessed based on corrected reward rates, and parameters related to the RL computational model were estimated accordingly. Subsequent analyses were conducted as follows. Firstly, the t-tests were performed to compare differences in behavioral outcomes and computational model parameters between low and high WM load conditions. Secondly, generalized linear mixed-effects models (GLMMs) were constructed to further analyze contributions from model-based and model-free strategies under varying WM load conditions, utilizing behavioral data that encompassed environmental information as well as participants' choices and obtained rewards during each trial. Thirdly, correlation analysis was employed to investigate the potential covariant relationships among measurements and parameters. Moreover, by fitting linear regression models, we aimed to explore how different variables collectively impacted task performance. Fourthly, mediation analysis sought to elucidate interaction mechanisms or indirect paths among variables. Detailed results of each analysis are presented below.

Comparative t-test analyses

Based on parameters estimated by the RL model along with each participant's corrected reward rate, t-tests were conducted to statistically validate the differences between low and high WM loads. For the behavioral performance, participants exhibited higher corrected reward rates in the low WM load condition than that in the high WM load condition ($t(49) = 11.2$, $p < 0.001$, $d = 1.58$), indicating that participants in the complex condition experienced lower reward acquisition rates and exhibited poorer overall performance (Fig. 1, Supplementary Tables 1 and 2). Furthermore, the inverse temperature in the low WM load condition was significantly higher than that observed in the high WM load condition ($t(49) = 9.99$, $p < 0.001$, $d = 1.41$), suggesting that participants' performance under the complex condition displayed a greater degree of randomness and a tendency towards exploration rather than exploitation. The learning rate in the low WM load condition was significantly higher than that in the high WM load condition ($t(49) = 7.88$, $p < 0.001$, $d = 1.11$), indicating a substantial reduction in information utilization during complex tasks. Additionally, eligibility trace was notably higher in the low WM load condition compared to that under the high WM loads ($t(49) = 2.51$, $p = 0.015$, $d = 0.36$), implying that the

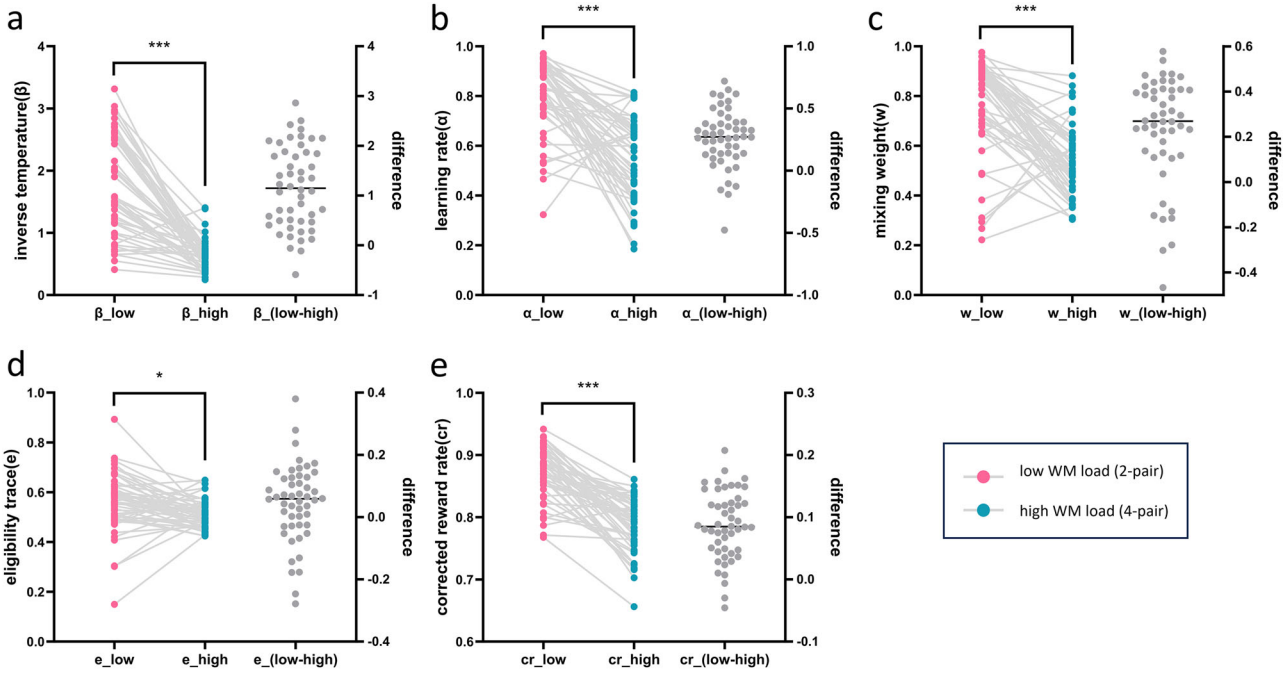


Fig. 1 | Paired t-test comparisons for reinforcement learning parameters and learning performance between high and low working memory load conditions. **a** The inverse temperature; **b** the learning rate; **c** the mixing weight; **d** the eligibility trace; and **e** the corrected reward rate. A single asterisk indicates a significant difference between the low WM load and the high WM load conditions, with $p < 0.05$, while three asterisks denote a significant difference with $p < 0.001$.

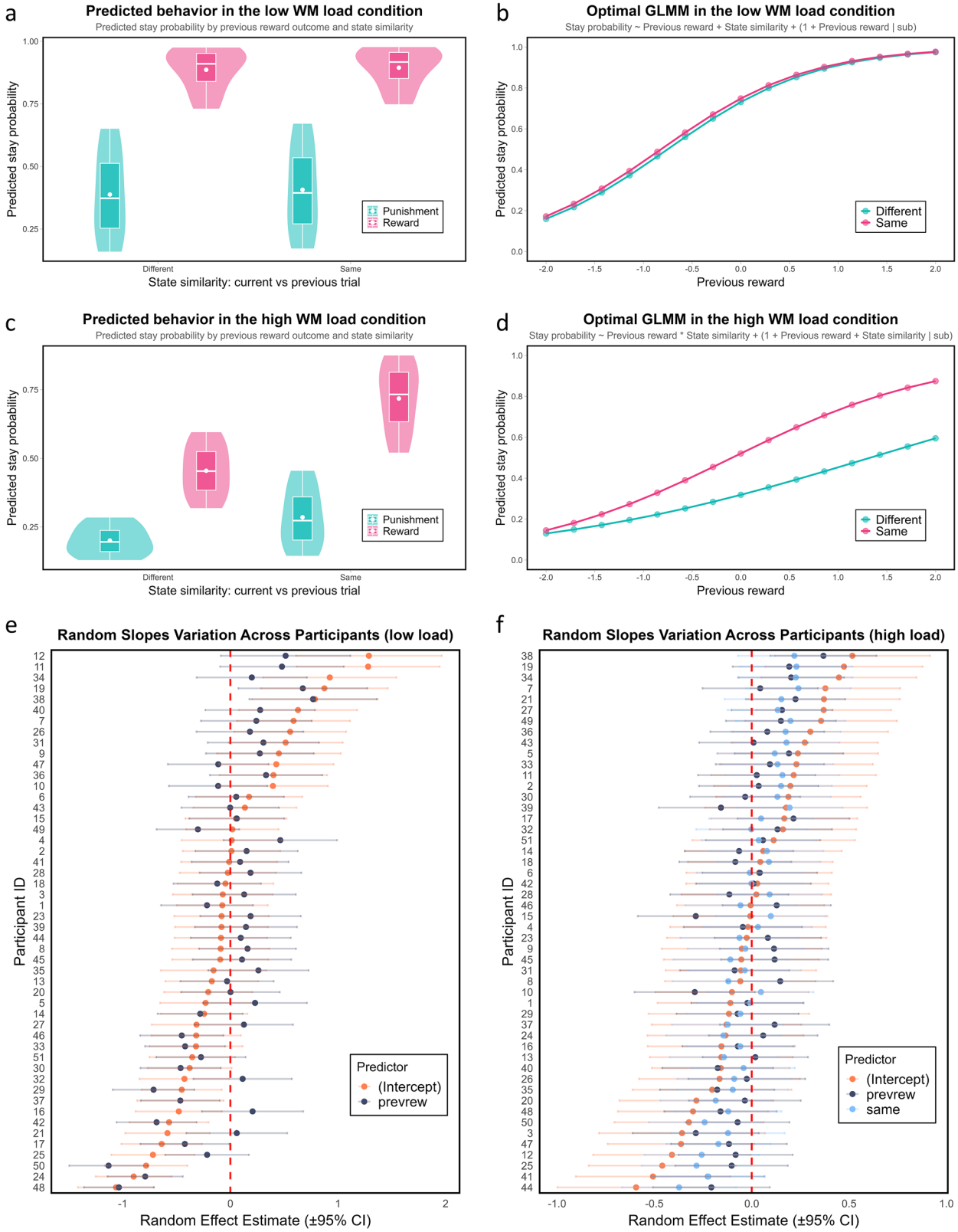
utilization of feedback outcome to update values of the first stage was considerably reduced within complex contexts. It should be noted that the mixing weight, which represented the degree of model-based control, was also significantly reduced in the high WM load condition compared to that observed in the low WM load condition ($t(49) = 6.95$, $p < 0.001$, $d = 0.98$; $W = 1140$, $p < 0.001$, $r = 0.79$). This suggests that in the complex contexts, participants were more inclined to employ simple and efficient model-free strategies rather than engaging model-based strategies which necessitated greater cognitive resources. These results demonstrate that under conditions of heightened WM demands, particularly when extensive information must be retained over prolonged delays, participants were unable to regulate their learning behaviors as effectively as they did within the low WM load scenarios.

GLMM analyses on different WM loads

To further explore the distinct learning strategies utilized under high and low WM loads, we fitted separate GLMMs for each condition. This approach enabled us to examine how participants' choices in the first stage and the rewards received in the second stage influenced their probability of staying with the same choice in the subsequent trial (see Supplementary Tables 3 and 5). The findings from these GLMMs corroborated previous t-test analyses by demonstrating that participants engaged in more model-based learning under low WM condition compared to high WM condition. To enhance clarity regarding variable interactions, we defined standardized previous rewards greater than 0 as "reward" and those less than 0 as "punishment".

In the optimal model for the low WM load condition, participants exhibited a pattern indicative of model-based learning. Previous rewards had a significant positive effect on stay probability ($\beta = 1.330$, $p < 0.001$), with no interaction reaching statistical significance (see Fig. 2a, b, Table 1). Random effects analysis revealed significant individual differences in both baseline stay probability (intercept variance = 0.333) and the impact of previous rewards (slope variance = 0.218), showing a moderate positive correlation between these factors ($r = 0.68$) (see Fig. 2e). The model fit indices of Akaike Information Criterion (AIC = 5644.46) and Bayesian Information Criterion (BIC = 5684.3) demonstrated strong explanatory power (see Supplementary Table 4). These results suggest that under low WM load condition, participants were more inclined to employ the model-based strategy, and their reward acquisition was not affected by state similarity. Furthermore, random slopes analysis highlighted crucial individual differences: participants displayed significant variability in their sensitivity to previous reward. The absence of significant random slopes for state similarity further confirmed the universality of model-based learning strategies under low WM loads.

Under the high WM load condition, participants demonstrated a pattern characteristic of model-free learning. The optimal model indicated significant main effects for both previous rewards ($\beta = 0.572$, $p < 0.001$) and state similarity ($\beta = 0.844$, $p < 0.001$), along with a significant interaction effect between them ($\beta = 0.357$, $p < 0.001$), indicating that the influence of previous reward was amplified when participants encountered identical states (see Fig. 2c, d, Table 2). Random effects analysis uncovered individual differences regarding baseline stay probability (intercept variance = 0.111) as well as the impacts associated with previous rewards (slope variance = 0.043) and state similarity (slope variance = 0.044). Moderate correlations were observed between baseline measures and previous reward effect ($r = 0.52$), while strong correlations emerged between baseline measures and state similarity effects ($r = 0.87$) (see Fig. 2f). The model fit indices (AIC = 5610.4, BIC = 5676.9) illustrated robust explanatory power (see Supplementary Table 6). These results imply that under high WM load condition, participants were more likely to adopt an effort-saving hybrid strategy or even rely exclusively on a pure model-free strategy. Their ability to acquire rewards was significantly affected by state similarity; they demonstrated heightened sensitivity to rewards and exhibited improved management of their selection behaviors when confronted with identical starting states. However, they no longer adhered to the principle of maximizing rewards when confronted with different starting states. The random effects unveiled crucial shifts in behavioral strategies: participants displayed individual differences in their sensitivity to both previous reward and state similarity, indicating varied weighting of hybrid strategies under WM constraints. Collectively, these random effects illustrate the "fragmentation" of learning strategies under high WM load condition; individuals employed hybrid strategies with



differing weights, yet overall, there was an increase in the weight assigned to the model-free strategy.

Correlation and linear regression analyses

To investigate the association among learning performance, RL parameters and intelligence under varying WM load conditions, we conducted correlation analyses (see Fig. 3). From the perspective of RL behavior, RL parameters were positively correlated under both low and high WM load conditions. For example, the inverse temperature exhibited a positive correlation with the learning rate (low load: $r = 0.63$, 95% CI [0.42, 0.77], $p < 0.001$; high load: $r = 0.56$, 95% CI [0.34, 0.73], $p < 0.001$), and the learning rate was positively correlated with the mixing weight (low load:

Fig. 2 | Generalized linear mixed-effects model (GLMM) results regarding model-free and model-based contributions to stay probability. a Predicted reinforcement learning (RL) behavior in the low WM load (2-pair) condition. The influence of previous rewards (either reward or punishment) on stay probability across two starting states (same or different from the preceding starting state) is comparable, reflecting characteristics of model-based learning. **b** Predicted results derived from the optimal model based on the influence of starting states in the low WM load condition. The trajectories representing same and different starting states are almost identical, further supporting evidence of model-based learning. **c** Predicted RL behavior in the high WM load (4-pair) condition. Previous rewards significantly affect stay probability across both starting states. Lower stay probabilities were observed in scenarios with a different starting state, indicative of tendencies associated with model-free learning. **d** Predicted results derived from the optimal model based on the influence of starting states in the high WM load condition. A significant increase in stay probability associated with previous rewards when originating from the same starting state; however, this trend demonstrates low sensitivity to previous rewards when considering different starting states, highlighting aspects of model-free learning. **e** Random effects at the participant-level regarding intercept and slope of previous reward (prevrew) in the low WM load condition. **f** Random effects at the participant-level encompassing intercept, slope of previous reward (prevrew), and state similarity (same) in the high WM load condition.

Table 1 | The optimal GLMM for the low WM load condition

Effect type	Term	Estimate	Std.Error	z-value	*p*-value	Variance (σ^2)	Std.Dev. (σ)	Corr
Fixed Effects	(Intercept)	1.001	0.090	11.070	<0.001	—	—	—
	prevrew	1.330	0.079	16.910	<0.001	—	—	—
	same	0.090	0.033	2.710	0.007	—	—	—
Random Effects	subnr (Intercept)	—	—	—	—	0.333	0.577	—
	subnr (prevrew)	—	—	—	—	0.218	0.467	0.680

Model: stay probability ~ previous reward (prevrew) + state similarity(same) + (1 + previous reward (prevrew) | sub).
Std.Error standard error, *Std.Dev* standard deviation, *Corr* correlation, *prevrew* previous reward, *same* state similarity, *subnr* subject number.

Table 2 | The optimal GLMM for the high WM load condition

Effect type	Term	Estimate	Std.Error	z-value	*p*-value	Variance (σ^2)	Std.Dev. (σ)	Corr
Fixed Effects	(Intercept)	−0.760	0.061	−12.449	<0.001	—	—	—
	prevrew	0.572	0.051	11.223	<0.001	—	—	—
	same	0.844	0.048	17.447	<0.001	—	—	—
	prevrew:same	0.357	0.042	8.562	<0.001	—	—	—
Random Effects	subnr (Intercept)	—	—	—	—	0.111	0.333	—
	subnr (prevrew)	—	—	—	—	0.043	0.208	0.520 (with Intercept)
	subnr (same)	—	—	—	—	0.044	0.209	0.870 (with Intercept), 0.220 (with prevrew)

Model: stay probability ~ previous reward (prevrew) * state similarity (same) + (1 + previous reward (prevrew) + state similarity (same)| sub).
Std.Error standard error, *Std.Dev* standard deviation, *Corr* correlation, *prevrew* previous reward, *same* state similarity, *subnr* subject number.

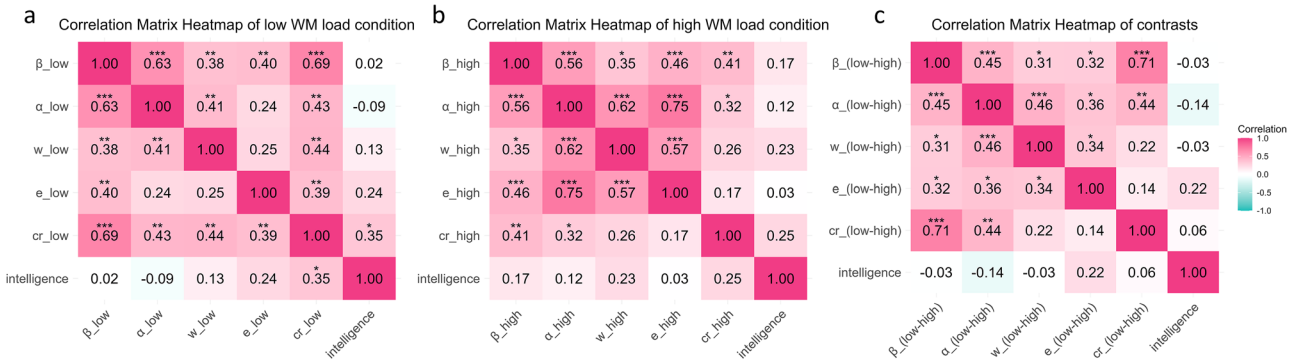


Fig. 3 | Correlation matrix heatmaps illustrating the relationships among reinforcement learning parameters, intelligence, and overall learning performance. **a** Correlation matrix heatmap corresponding to the low WM load condition. **b** Correlation matrix heatmap corresponding to the high WM load condition. **c** Correlation matrix heatmap depicting contrasts: differences between the low WM load condition and the high WM load condition. β represents the inverse temperature. α denotes the learning rate. *w* signifies the mixing weight. *e* stands for the eligibility trace. *cr* refers the corrected reward rate, representing overall learning performance.

r = 0.41, 95% CI [0.15, 0.62], $p = 0.003$; high load: r = 0.62, 95% CI [0.41, 0.76], $p < 0.001$). These findings suggest that model-based agents effectively extracted experience from previous trials through information exploitation and value updating. In terms of learning performance, the corrected reward rate was significantly correlated with RL parameters, especially with the inverse temperature (r = 0.69, 95% CI [0.51, 0.81], $p < 0.001$) and the mixing weight (r = 0.44, 95% CI [0.19, 0.64], $p = 0.001$) in the low WM load condition, suggesting that model-based agents gained more rewards in this

context. And in the high WM load condition, significant correlations emerged between the corrected reward rate and both the inverse temperature (r = 0.41, 95% CI [0.14, 0.61], $p = 0.003$) and the learning rate (r = 0.32, 95% CI [0.04, 0.55], $p = 0.026$); this suggests that participants who were proficient at exploitation and information utilization achieved higher rewards under this circumstance. Moreover, intelligence demonstrated a significant correlation with the corrected reward rate in the low WM load condition (r = 0.35, 95% CI [0.07, 0.57], $p = 0.013$). However, no significant correlation was observed between learning performance and intelligence in the high WM load condition.

To further explore the associations among WM load, participants' learning performance, RL parameters, and intelligence, we performed linear regression analyses. Our primary objective was to examine how each variable related to learning performance, while controlling for other factors. The mixing weight parameter served as an indicator of model-based RL. We encoded WM load as 1 for low load condition and 2 for high load condition accordingly. Considering the reliability and rationality of the model, from the full model to all stepwise regression models, we selected an optimal model where the corrected reward rate was determined a function of WM load, the mixing weight, and intelligence (intercept = 0.78, SE = 0.04, t = 17.57, $p < 0.001$; $R^2 = 0.60$, see Fig. 4). Additionally, we referenced the AIC and BIC values to further refine our model selection process (AIC = −357, BIC = −344), ensuring that the chosen model provided the best fit for the data while maintaining an appropriate balance between goodness-of-fit and complexity. Our findings indicated that WM load negatively predicted task performance (β = −0.07, SE = 0.01, t = −7.17, $p < 0.001$), suggesting that higher WM loads hindered participants' ability to collect rewards effectively. Intelligence scores positively predicted task performance (β = 0.003, SE = 0.001, t = 2.62, $p = 0.01$), suggesting that participants with higher intelligence were more adept at obtaining rewards in the RL task. Additionally, the mixing weight was found to positively predict learning performance (β = 0.08, SE = 0.02, t = 3.41, $p < 0.001$). In general, participants achieved greater rewards by employing model-based strategies, and the use of such strategies was jointly influenced by fluid intelligence and task complexity (see Table 3).

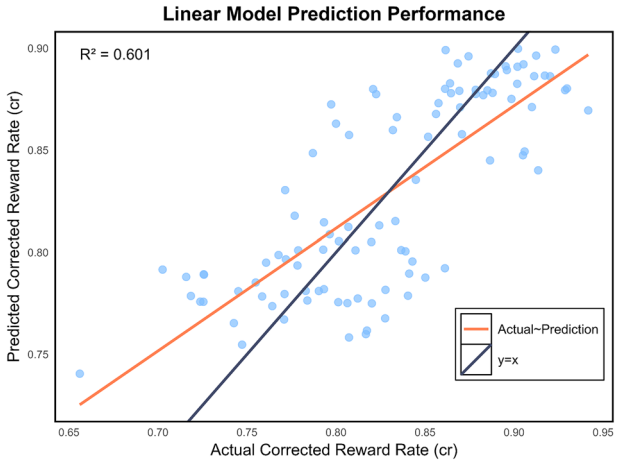


Fig. 4 | Prediction performance of optimal linear model (Corrected reward ~ Mixing weight + WM load + Intelligence). The scatter plot comparing actual versus predicted values clusters closely around the reference line (y = x), with an R^2 value of 0.601, indicating strong predictive accuracy.

The mediation model

We conducted an exploratory mediation analysis to test whether the effect of WM load on task performance (corrected reward rate) was mediated by the mixing weight of model-based strategies. Bootstrapping results based on 1000 samples revealed that the impact of WM load on corrected reward rate was partially mediated by mixing weight of model-based learning strategies (see Fig. 5, Table 4). WM load negatively predicted mixing weight (path a = −0.232, $p < 0.001$), which in turn positively predicted reward rate (path b = 0.090, $p < 0.001$). The indirect effect (ACME = −0.021, 95% CI [−0.037, −0.008], $p = 0.002$) accounted for 24% of the total effect (Proportion Mediated), while the direct effect (ADE = −0.066, 95% CI [−0.088, −0.044], $p < 0.001$) remained significant. This suggests that higher WM load reduced the reliance on model-based strategies, thereby impairing reward acquisition, though additional unmediated pathways also contributed to the total effect (Total = −0.087, $p < 0.001$).

Considering the limited sample size, we implemented a Monte Carlo simulation with 1000 replications using the observed path coefficients (a and b paths) alongside our sample size ($N = 100$; comprising 50 participants under both conditions). This approach enabled us to determine what percentage of these simulated analyses would yield statistically significant mediation effects of Average Causal Mediation Effects (ACME) at α = 0.05. This process allowed us to empirically estimate our study's power to detect true mediation effects[44]. The results demonstrated that the current mediation analysis achieved a statistical power of 96.4% (α = 0.05) for detecting the ACME, given the specified model parameters and sample size. The outcomes of the Monte Carlo simulations indicated our analysis maintained reasonable power to identify mediation effects within the constraints imposed by the present sample size.

Discussion

The current study investigated the interplay between mental abilities and model-based versus model-free learning strategies, as well as how the associations affected learning performance. Our findings indicated that participants flexibly adopted different learning strategies depending on varying WM load conditions which necessitated differing levels of mental effort. Under high WM load, where greater mental effort and more cognitive resources were required, participants tended to employ a model-free learning strategy. Conversely, under low WM load, where less mental effort was needed, participants were more inclined to adopt a model-based learning strategy. Furthermore, it was found that the model-based learning strategy mediated the impact of WM loads on RL performance. While participants' intelligence scores showed a significant association with RL performance outcomes, they did not correlate with model-based or model-free strategies in the current study.

Table 3 | Linear regression coefficients indicating the effects of mixing weight, working memory load and intelligence on corrected reward rate

Predictor	Estimate	Std.Error	t-value	*p*-value	Std. Estimate	95% Confidence interval	
						Lower	Upper
Intercept	0.777	0.044	17.570	<0.001			
Mixing weight	0.080	0.023	3.410	<0.001	0.27	0.112	0.427
WM load	−0.069	0.010	−7.170	<0.001	−0.562	−0.717	−0.406
Intelligence	0.003	0.001	2.620	0.01	0.172	0.042	0.301

Std.Error standard error, *Std.Estimate* standard estimate, *WM* working memory.

Table 4 | Results of mediation effect analysis

Effect measure	Estimate	95% Confidence interval		*p*-value
		Lower	Upper	
ACME	−0.021	−0.037	−0.008	0.002
ADE	−0.066	−0.088	−0.044	<0.001
Total Effect	−0.087	−0.104	−0.071	<0.001
Prop. Mediated	0.24	0.088	0.437	0.002

ACME Average Causal Mediation Effect, *ADE* Average Direct Effect, *Prop. Mediated* Proportion Mediated.

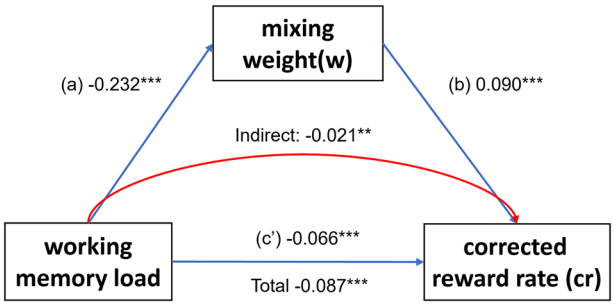


Fig. 5 | Mediation analysis of working memory load on corrected reward rate with mixing weight as a mediator. The mediation analysis revealed a significant direct negative effect of working memory load on corrected reward rate. In the mediated pathway, working memory load showed a significant negative effect on mixing weight, which in turn had a significant positive effect on corrected reward rate, resulting in a significant indirect effect. The total effect of working memory load on corrected reward rate was also significant.

The present study indicated that WM load served as a crucial role in human arbitration between model-based and model-free RL processes. In the low WM load condition, participants devoted less mental effort to the memorization of stimulus-action-outcome associations and retained more cognitive resources for implementing flexible and prospective model-based learning. Conversely, under high WM load condition, participants were more likely to implement a hybrid or pure model-free strategy that was resource-efficient and immediate. Under high WM loads, more mental effort was required to memorize the stimulus-action-outcome associations, and fewer resources remained available for the RL process. These findings suggest that human brains exhibit flexibility in balancing benefits against mental effort throughout the task; they estimate the expected rewards associated with each strategy governed by distinct systems before weighing these rewards against the respective costs[28]. It is also important to note that the manipulation of increasing WM load by incorporating additional environmental structures may elevate overall task difficulty. A more complex transition structure could heighten individuals' planning challenges by necessitating simultaneous monitoring of multiple potential planning paths —this cognitive demand aligns with Kool et al.'s findings, which suggest that task planning complexity can influence the task-specific cost of mental effort[45], thereby biasing strategy selection and impairing learning performance[15,46–52]. Furthermore, the increased complexity and uncertainty associated with transition structures in more challenging conditions may correlate with higher intrinsic effort costs and constrain the availability of cognitive control[36,43,49,53,54]. The deficits observed in model-based learning under high WM load conditions might be linked to impaired exertion of cognitive control[49,54,55].

The random effects derived from optimal GLMMs under low and high WM load conditions revealed distinct correlations. In the low WM load condition, participants exhibiting higher baseline stay probabilities demonstrated greater sensitivity to previous reward, indicating that this "cautious-optimizing" covariation pattern may reflect trait-like learning strategy preferences. In the high WM load condition, a strong correlation between baseline stay probability and state similarity effect suggests that conservative learners relied more heavily on state-matching heuristics. Meanwhile, the moderate correlation between baseline-reward sensitivity and individual specificity in reward-driven behavior was preserved. These patterns of individual differences suggest that during RL processes, both learning performance and strategic arbitration may be influenced by personality traits or emotion states.

No significant correlation was observed between intelligence and mixing weight. This finding aligns with studies[29] that employed Kool et al.'s[27] novel two-stage task, but contrasts with research[23] utilizing the traditional two-stage task developed by Daw and his colleagues[9]. In Kool et al.'s novel task, deterministic transitions and faster drifting reward rates encouraged participants to adopt the model-based strategy; a model-based agent could leverage the internal model of the task structure to maximize rewards. Therefore, task performance and learning strategies relied more on task-based WM capacity to construct the task structure. On the other vein, the traditional two-stage task presented greater complexity due to its higher randomness; thus, general intelligence was highly desired to account for the probabilistic transition. It can be inferred that the relationship between model-based strategy adoption and intelligence is unstable and affected by characteristics inherent in different two-stage paradigms.

In examining the factors influencing learning performance, it was found that inverse temperature positively correlated with corrected reward rates under both high and low WM loads. This finding suggests that selecting actions based on exploitation of known information was associated with higher rates of obtained rewards. Under low WM loads, corrected reward rates were significantly associated with both model-based mixing weight and intelligence scores; however, these correlations diminished under high WM loads. One possible explanation is that the complexity associated with high WM demands may have impeded participants' ability to establish accurate cognitive models[22], thereby preventing them from maximizing rewards through a more model-based strategy. This speculation is partially corroborated by participants' feedback collected after task completion; several participants reported incorrect transition structures. Notably, the relationship between intelligence and learning performance varied across different WM load conditions. In the low WM load condition, intelligence exhibited a stronger positive correlation with task performance. In contrast, in the high WM load condition, where intensive utilization of WM capacity was required, the association between intelligence and performance was attenuated. This may be attributed to the fact that task-specific cognitive resources became more critical than general intellectual ability within such contexts. Further investigation is warranted by designing an additional WM task alongside the two-choice RL task, which could further elucidate the relationships among WM, intelligence and RL processes through independent manipulation of WM demands.

Furthermore, to gain a deeper understanding of how WM load, model-based strategy, and human intelligence collectively contribute to learning performance, both linear regression models and mediation analyses were conducted to provide additional insights. The linear regression analysis indicated that WM loads negatively predicted participants' learning performance; specifically, higher WM loads were associated with lower reward rates. Elevated WM load may hinder participants from allocating sufficient cognitive resources to balance the memorization of task structure and the interpretation of outcome, ultimately impairing their ability to secure greater rewards. Simultaneously, both model-based implementation and intelligence positively predicted participants' performance in selecting rewarded options. A greater application of model-based strategy facilitated a better understanding of task structures and enabled more accurate predictions regarding action values, leading to an increase in rewarded choices. Higher levels of intelligence correlated with enhanced reasoning capabilities and flexibility in understanding task structures as well as stimulus-action-outcome associations, which resulted in improved choice outcomes. More importantly, the mediation analysis further revealed that the mixing weight representing the contribution of model-based learning served as a partial

mediator in the association between WM load and final learning performance. This finding suggests that model-based learning strategy may mediate the impact of WM load on learning outcomes. Specifically, engaging in tasks under high WM load inevitably requires greater mental effort and cognitive resources compared to low WM load condition; this increased demand impairs participants' initiative to employ model-based strategy and subsequently hinders their performance in reward collection. However, given the relatively small sample size, the results from the mediation analysis may have a certain degree of uncertainty, and thus should be interpreted and used with caution.

Several limitations are present in the current study. First, the generalizability of our findings is constrained by the relatively small sample size. The limited sample may have reduced statistical power and restricted the applicability of certain analytical approaches, particularly mediation analysis. While observed patterns are theoretically meaningful, their robustness should be verified through larger-scale replications. Second, the current design confounded WM load with general task difficulty; this may limit specificity regarding conclusions about the role of WM itself. High WM load might indirectly increase overall task difficulty, resulting in performance differences that reflect global cognitive demands rather than effects specific to WM. Future research can further investigate these effects by incorporating an explicit WM task while participants engage in the two-stage RL task. By independently manipulating WM demands, it may be possible to determine participants' threshold for transitioning from a model-based to a model-free strategy. It will be valuable to further explore how WM and intelligence influence learning performances under these circumstances.

In summary, this study demonstrates that individuals can flexibly adjust their learning strategies based on varying circumstances associated with different levels of WM loads. The utilization of model-based strategy, together with WM capacity and intelligence, influences an individual's learning performance. Furthermore, the impact of WM load on an individual's learning performance may be mediated by the use of model-based learning strategy. The current study sheds light on humans' flexible mechanism for adjusting algorithms across diverse strategies according to specific task requirements, thereby enhancing both adaptability and efficiency during RL processes.

Methods

Participants

All participants were recruited from local universities through online platforms in Beijing. In accordance with the Declaration of Helsinki, the experimental protocols were approved by the ethic committee of Institute of Psychology, Chinese Academy of Sciences.

We utilized G*power to determine the required sample size for our study. In alignment with our primary hypothesis, we opted for matched pairs t-tests as the statistical method, anticipating an effect size of 0.5 ($\alpha = 0.05$, $1-\beta = 0.95$, one-tailed, $\eta^2 = 0.5$). The analysis indicated that a minimum sample size of 45 participants was necessary. Consequently, we recruited a total of 50 human participants (23 males and 27 females). The age range of the participants spanned from 18 to 29 years, with a mean age of 21.38 years (SD = 3.50). All participants had normal or corrected-to-normal vision and exhibited no difficulties in color discrimination. Additionally, none had a family history of mental illness, and all were right-handed. Informed consent was obtained from each participant prior to the involvement in the study. Upon completion of the experiment, each participant received basic remuneration along with performance-based rewards during the task; these payments ranged from ¥60 to ¥80 in total.

Procedure

Participants completed the two-stage RL tasks on a computer, with stimulus presentation and behavioral response recording implemented through PsychoPy (version 2024.1.5). Participants' fluid intelligence was assessed through Cattell's Culture Fair Intelligence Test[38,39] which comprises four independent sections: serialization (12 items), classification (14 items), reasoning (12 items), and topology (8 items). The number of correctly answered items within a limited time frame served as an indicator of fluid intelligence. The methods and analyses employed in the present study were not preregistered.

The current paradigm was adapted from Kool et al.'s[27] two-step RL task, incorporating both a low WM load condition [similar to Kool et al.'s task] and a high WM load condition (Fig. 6a). These two conditions were organized into separate blocks, with their order balanced across participants; the participants were informed about the sequence at the beginning of the task. In the low WM load condition (2-pair), there were two fixed pairs of doors presented in the first stage, and two distinct bears were shown in the second stage (Fig. 6b), mirroring the original design established by Kool and his colleagues. The door pairs remained constant throughout the experiment; for example, one pair consistently consisted of a red door and a green door, while another pair comprised yellow and blue doors. The positions of these doors on either side of the screen were balanced across trials. In the high WM load condition, four fixed pairs of doors appeared in the first stage alongside four different bears displayed in the second stage (Fig. 6d). In both conditions, the experimental background was set to white; each door was approximately 287 pixels by 452 pixels while each bear was around 363 pixels by 450 pixels. The visual angle was approximately set at 60°.

Each trial consisted of two stages under both low and high WM load conditions. In the first stage, one pair of doors was presented on the screen, and participants were instructed to choose one door. Each selected door deterministically led to encountering one specific bear in the second stage (Fig. 6b, d). Participants had a time limit of 2 s within which to make their selection by pressing the "F" key for left-door selection and "J" key for right-door selection. Upon selection, a red frame highlighted their chosen door; this selection process took a total duration of 3 s. Subsequently, in the second stage, an image of one bear was displayed centrally on the screen. Participants were required to request strawberries from the bear by pressing the "SPACE" key within 2 s. Following this action, feedback regarding the number of strawberries obtained during that trial was presented on the screen for 1 s. The quantity of strawberries provided by each bear adhered to an independent Gaussian random walk distribution ($\mu = 0$, $\sigma = 2$), with values ranging from 0 to 9 (Fig. 6c). The intertrial interval was set at 2 s. In conditions characterized by low WM load, each pair of doors was presented a total of 60 times; in high WM load conditions, each pair appeared 30 times. Participants were given a one-minute break after every 40 trials within each condition, and there was a three-minute break between different conditions. Prior to commencing the formal task, all participants underwent a practice session consisting of 20 trials to familiarize themselves with the task procedure. Each condition lasted about 18 min.

In addition to standard trials, four probe trials were administered at intervals corresponding to the 64th, 79th, 94th, and 109th trials for each condition in order to assess whether participants had accurately learned the mapping between doors and bears. During each probe trial, a prompt was presented for 2 s, and participants were instructed to identify which bear could provide substantial strawberries by choosing one door from the pair of doors (the pair configuration mirrored that used in standard trials) (Fig. 6e). Participants were required to recall previously learned transition structures associated with tasks while attempting to select doors linked with bears capable of providing more rewards. Correct selections resulted in an award of one hundred strawberries, whereas incorrect selections yielded no rewards during these probe trials. Analysis on the numbers of rewards collected from these probes provided insights into when participants acquired knowledge regarding transition structures alongside stimulus-action-outcome associations.

The experimenter provided comprehensive instructions to the participants, who were permitted to review the instruction document and pose questions as necessary. Through familiarization with these instructions, participants were required to grasp several key points: first, the independent fluctuation of strawberry availability when interacting with different bears; second, the deterministic transitions between doors and bears, where each door consistently led to a specific bear; third, the procedure for selecting doors and collecting strawberries from bears. Participants were explicitly informed

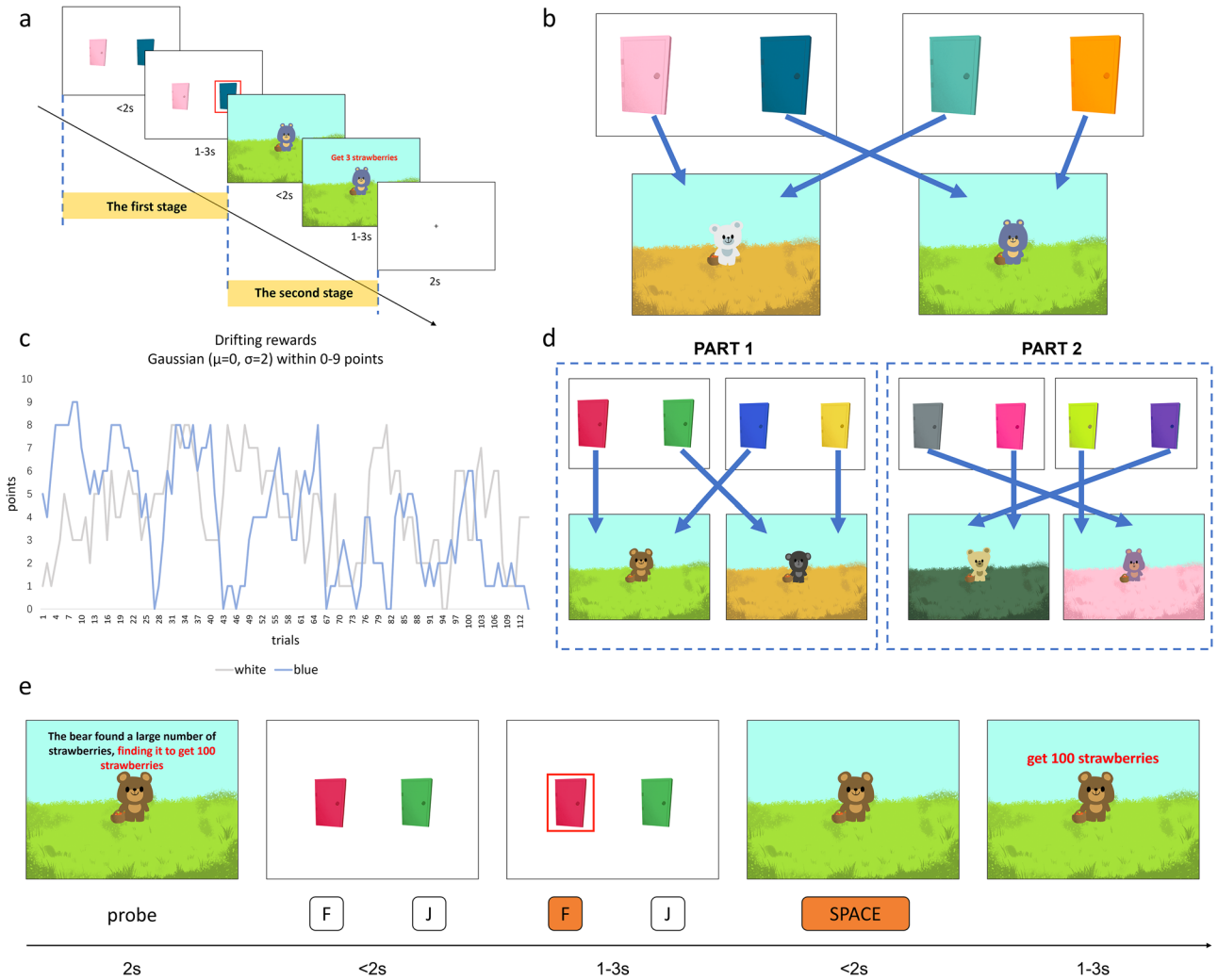


Fig. 6 | Two-stage reinforcement learning tasks. a Each trial consists of two stages. In the first stage, participants were required to choose either the right or left door to enter; subsequently, in the second stage, participants collected strawberries from a bear and received feedback regarding their choices. **b** Transition structure for the low WM load condition. The pink and cyan doors deterministically led to the white bear, while the dark blue and orange doors deterministically led to the blue bear. **c** Random fluctuations in rewards. The number of strawberries that participants received from each bear ranged from 0 to 9, following an independent Gaussian random walk ($\mu = 0$, $\sigma = 2$). **d** Transition structure for the high WM load condition. The red and blue doors deterministically led to the brown bear, whereas the green and yellow doors deterministically led to the black bear. The rose and purple doors deterministically led to the buff bear, and the gray and lime doors deterministically led to the purple bear. **e** During each probe trial, participants were instructed to identify which bear could provide substantial strawberries by choosing one door from the pair of doors.

that their goal was to collect as many strawberries as possible from the bears across both conditions. Through repeated attempts, they could discover that the number of strawberries given by different bears varied within a certain period of time; thus, participants had to decide which door to choose in the first stage in order to encounter the bear that could offer more strawberries. Given that strawberry availability was constantly changing, participants needed to continuously monitor feedback fluctuations in the second stage while striving for optimal choices made during the first stage. Moreover, it is important to note that the amount of strawberries collected directly influenced participants' monetary compensation associated with their performance on this task. For every 250 strawberries gathered, participants would receive a payment of ¥5, and the total number of strawberries obtained would be displayed on the final trial of each condition.

Task performance measurement and RL model parameter estimation

Preprocessing was conducted using R 4.4.1, encompassing steps such as deleting missing trials and encoding participants' behaviors.

Behavioral performance was measured as follows. The number of strawberries generated for each participant was independent, which also meant that the maximum number of strawberries they could obtain was independent. Therefore, we used a corrected reward rate to evaluate their task performance by computing the ratio of the actual number of strawberries obtained to the theoretical maximum number of strawberries (see Eq. (1)). The theoretical maximum number was defined as the sum of all higher-value rewards that could be obtained by consistently choosing the more rewarded door-bear associations at each trial timepoint. The actual number of strawberries obtained reflected the extent to which participants made optimal choices throughout the experiment. Higher corrected reward rates represented better performance. This measurement was used for subsequent statistical analyses to compare participants' different task performance across low and high WM load conditions.

$$corrected\ reward\ rate\ (cr) = \frac{\text{actual obtained quantity}}{\text{maximum obtainable quantity}} \quad (1)$$

The dual-systems RL computational model was built. We employed a well-established and validated dual-system RL model[2,9,12,27,28] to estimate parameters reflecting model-based and model-free contributions to decision-making behaviors. The packages and foundational model framework utilized in our analysis were derived from prior studies[27,28]. The model comprised both a model-free system and a model-based system, which differed in their levels of environmental representation as well as the methodologies used to estimate action and values in certain states. These values of state-action pairs were denoted as the function Q(s,a)[5,9,56,57]. The model-free system simply updated action values based on reward prediction error (RPE) comparisons between actual outcomes and expected outcomes. When an action resulted in a more favorable outcome than anticipated, the value of this action increased; conversely, if the outcome was less favorable than expected, its value decreased. In contrast, the model-based system updated values according to not only feedback but also an internal model of the environment to formulate plans that incorporated both environmental transition structure and reward functions, gradually progressing towards the final goals. Throughout the entire learning process, transition structure and reward functions were continuously maintained and updated in real-time to guide decisions by considering potential consequences of actions. In RL tasks, distinguishing between the behaviors of model-free and model-based agents was justified by the latter's ability to utilize generalized information. For instance, when a participant employing a model-based strategy chose the pink door and encountered the white bear that provided nine strawberries, he/she would be more likely to choose the pink door as well as the cyan door in subsequent trials, rather than limiting their choices solely to the pink door. In short, while model-based agents could establish connections between doors leading to bears, model-free agents relied exclusively on rewards[26,27].

According to previous studies[28], we estimated and analyzed the following modeling parameters to elucidate the differences in learning processes between low and high WM load conditions. The computational model simulated participants' learning process by incorporating behavioral data (choices and obtained rewards) and environmental information (transition structures and available options) from each trial. The value associated with each action (selection behavior) was updated on a trial-by-trial basis, relying on calculations of RPE and state prediction error (SPE). The resulting parameters reflected various learning characteristics, including degree of information utilization, tendencies toward exploration versus exploitation, and preferences for model-based versus model-free strategies.

For computing model-free strategy, model-free agents update values of state-action pairs $Q(s, a)$ at stage i ($i = 1, 2$) and trial t in accordance with the SARSA (state-action-reward-state-action) temporal difference learning algorithm[58] in the two-stage task. The update rule is expressed as follows:

$$Q_{MF}(s_{1,t}, a_{1,t}) \leftarrow Q_{MF}(s_{1,t}, a_{1,t}) + \alpha\delta_{1,t} + \alpha\delta_{2,t}e_{2,t}(s_{2,t}, a_{2,t}) \quad (2)$$

$$Q_{MF}(s_{2,t}, a_{2,t}) \leftarrow Q_{MF}(s_{2,t}, a_{2,t}) + \alpha\delta_{2,t} \quad (3)$$

where

$$\delta_{i,t} = r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MF}(s_{i,t}, a_{i,t}) \quad (4)$$

is the RPE that reflects the discrepancy between the actual value and the expected value at the current stage. The actual value comprises the reward received during this stage and the expected value for subsequent stage. The effect of RPE is jointly determined by three factors: learning rate (α), reward prediction error (δ) and eligibility trace (e). The learning rate is associated with the efficiency of value-updating and represents how much new information will be integrated into forming expectations for future rewards, ranging from 0 to 1. A higher learning rate signifies more efficient updates of action values based on outcomes, suggesting that agents can incorporate more information from experience.

The eligibility trace (e) captures how outcomes in the second stage affect action values in the first stage. The eligibility trace ($e_{i,t}(s_{i,t}, a_{i,t})$) is initialized to 0 at the beginning of each trial and is updated prior to Q-value updating:

$$e_{i,t}(s_{i,t}, a_{i,t}) = e_{i-1,t}(s_{i,t}, a_{i,t}) + 1,\ i = 1 \quad (5)$$

representing the extent to which the second-stage RPE affects updates of reward expectations for the first-stage actions. Then the eligibilities of all state-action pairs are decayed by λ (ranging from 0 to 1) after the updating of first stage.

$$e_{i,t}(s_{i,t}, a_{i,t}) = e_{i-1,t}(s_{i,t}, a_{i,t})(1 - \lambda),\ i = 2 \quad (6)$$

It should be noted that in the first stage, the reward is always equal to 0 since no rewards are obtained; thus, RPEs in the first stage solely reflect differences between expected values across both stages.

$$\delta_{1,t} = Q_{MF}(s_{2,t}, a_{2,t}) - Q_{MF}(s_{1,t}, a_{1,t}) \quad (7)$$

However, in the second stage, participants receive feedback regarding the rewards. Given that there is no next stage, RPE is driven by the discrepancies between received reward values and expected values.

$$\delta_{2,t} = r_{2,t} - Q_{MF}(s_{2,t}, a_{2,t}) \quad (8)$$

The action values for both first and second stages are updated at the second stage, as the updating of the first stage relies on the eligibility trace parameter generated in the second phase.

To measure model-based strategy, model-based agents acquire knowledge about task transition structures that map state-action pairs from the first stage to subsequent states in the second stage. They integrate these task structures with feedback rewards—model-free values obtained directly during the second stage—to compute cumulative state-action values throughout the entire learning process via continuous iteration. As described above, model-based agents acquire feedback rewards in the second stage in the same manner as model-free agents. This is because, for both strategies, the estimated values in the second stage depend on prior feedback rather than on the task structure itself. Model-based values are articulated through Bellman's equation[59], wherein state-action pairs values are determined based on all possible behavioral outcomes and all transition structures are assumed to be known. For example, within our experiment, there are two actions—a_{1A} and a_{1B}—in the first stage; each action deterministically leads to a specific subsequent state (s_{2A} or s_{2B}) in the second stage, where only action a_2 can be taken. According to Bellman's equation:

$$Q_{MB}(s_1, a_j) = P(s_{2A}|s_1, a_j)\max_a Q_{MF}(s_{2A}, a) + P(s_{2B}|s_1, a_j)\max_a Q_{MF}(s_{2B}, a), \quad (9)$$

where $P(s_i|s_{i-1}, a_j)$ is the transition probability from the current state (s_{i-1}) to the subsequent state (s_i) by taking action (a_j). Based on the deterministic transition structure in our learning tasks, the probabilities are always either 0 or 1. If a model-based agent has learned the correct transition structure in the task, he/she will compute each state-action value as follows:

$$Q_{MB}(s_1, a_{1A}) = P(s_{2A}|s_1, a_{1A})\max_a Q_{MF}(s_{2A}, a) = Q_{MF}(s_{2A}, a_2) \quad (10)$$

$$Q_{MB}(s_1, a_{1B}) = P(s_{2B}|s_1, a_{1B})\max_a Q_{MF}(s_{2B}, a) = Q_{MF}(s_{2B}, a_2) \quad (11)$$

These model-based estimated values are recomputed at each trial by using the updated information of transition structure and rewards in the second stage.

To infer the contributions of model-free versus model-based strategies in learning, we computed the mixing weight (w). Rather than relying solely on either a single model-free or model-based strategy, participants typically

employ a hybrid model that integrates both strategies according to:

$$Q_{net}\left(s_{i,t}, a_{j,t}\right) = wQ_{MB}\left(s_{i,t}, a_{j,t}\right) + (1-w)Q_{MF}\left(s_{i,t}, a_{j,t}\right), \quad (12)$$

where (w) represents the contribution of model-based strategy, while ($1-w$) signifies that of the model-free strategy. A mixing weight value approaching 1 indicates a strong reliance on the model-based strategy, whereas a value nearing 0 reflects a predominant use of model-free strategy. After computing the values for state-action pairs, the softmax method is used to convert these values into probabilities for each action in a specific state. The probability of choosing action (a) on trial (t) at state (s) is computed as follows:

$$P(a_{i,t} = a|s_{i,t}) = \frac{exp\left(\beta Q_{net}\left(s_{i,t}, a\right)\right)}{\sum_{a'} exp\left(\beta Q_{net}\left(s_{i,t}, a'\right)\right)} \quad (13)$$

where (a') indexes all currently available actions and (β) denotes the inverse temperature parameter. It is assumed that the probability of taking an action is determined by the proportion of its value relative to all action values, and the exploitation of the value information is influenced by the inverse temperature. The inverse temperature (β) governs the exploration-exploitation trade-off, namely the randomness of the choice. Higher values of inverse temperature correspond to greater exploitation of known information when selecting actions, while lower values indicate greater exploration tendencies. As this parameter increases, participants demonstrate enhanced exploitation of learned information, preferentially selecting actions associated with higher expected values. Conversely, as the parameter approaches zero, action selection becomes less sensitive to value differences, resulting in uniformly distributed choice probabilities.

Additionally, we set priors for parameters to facilitate initial fitting, and these parameters were adjusted through constant iteration to best align with participants behaviors. In subsequent statistical analyses, computational model parameters (the mixing weight, the inverse temperature, the learning rate, and the eligibility trace) were compared between low WM load (2-pair) and high WM load (4-pair) conditions. These parameters were also utilized for correlation analyses and linear modeling as well as mediation analyses.

Statistical analyses

We conducted several analyses on the current data. First, one-tailed t-tests were performed to compare differences in behavioral performance (corrected reward rate) and computational model parameters (mixing weight, learning rate, etc.) between low and high WM load conditions.

Second, we constructed GLMMs using R packages to analyze the contributions of model-based and model-free strategies to learning performance under low and high WM load conditions, as informed by prior research[27,30]. Participants' choice and reward information (including the number of obtained rewards) from each trial were used to fit models separately for each WM load condition. The best-fitting model was selected based on the corrected AIC (AICc). Model selection and result visualization were carried out using the AICcmodAvg package[60] alongside the ggeffects package[61]. According to two different optimal models corresponding to each condition, we compared the effects of previous rewards (prevrew), differences in provided rewards from previous trials (prevrewdiff) and state similarity of options in the first stage between current and previous trials (same) on the probability of revisiting a door (stay probability). This analysis reflects participants' learning differences based on their behavioral patterns. Specifically, a main effect of previous rewards without interaction can be interpreted as a component of a model-based strategy, suggesting that participants can maximize rewards regardless of whether the starting state is the same or different. This is a sign that the action value is generalized based on environmental information. Conversely, a main effect of previous reward with an interaction between previous rewards and state similarity can be viewed as a component of model-free strategy[27]. Participants are unable to follow the principle of maximizing rewards when confronted with different starting states. This inability to generalize information based on environmental clues is a manifestation of model-free learning[27,30]. In each model, we incorporated participant-level random intercepts to account for individual differences in baseline stay probability. Previous studies have demonstrated that individuals also varied in their strategy selection based on reward stakes and environmental characteristics[30,45], which may influence their stay probabilities. Therefore, building upon optimal random-intercept models, we further integrated random slopes for previous reward and state similarity to assess whether these effects varied across participants.

Third, the correlation analyses were used to preliminarily investigate the potential covariation among measurements and parameters, serving as preparation for the linear regression model. Given that not all data obeyed a normal distribution, both parametric (Student's t-test, Pearson correlation) and non-parametric (Wilcoxon W test, Spearman correlation) approaches were applied where appropriate. Moreover, by fitting the linear regression models utilizing the lme4 package[62], we aimed to explore how different variables impacted task performance. These models assessed how the corrected reward rate was predicted by contributions from model-based strategy, WM load, and intelligence scores.

Finally, the mediation analysis aimed to elucidate the interaction mechanisms or indirect paths among variables. Based on the potential associations identified among these variables, we proposed two mediation models: first, the mixing weight may mediate the effect of WM load on learning performance; second, the mixing weight could mediate the contribution of intelligence on learning performance. Given the limited sample size, we conducted tests of statistical power by simulating 1000 datasets while repeatedly performing mediation analyses and calculating the proportion of significant ACME results. We employed Monte Carlo simulation methods utilizing data resampling techniques to comprehensively evaluate joint effects within mediation analysis[44]. By directly simulating real research scenarios, this approach yields highly consistent results with actual data analysis outcomes, making it particularly suitable for testing complex mediation models.

Data availability

The datasets generated and analyzed during the current study are available in the Open Science Framework (OSF) repository: https://osf.io/t2vdz/files/osfstorage.

Code availability

The underlying code for this study is available in OSF repository and can be accessed via this link: https://osf.io/t2vdz/files/osfstorage.

Received: 24 April 2025; Accepted: 15 September 2025;
Published online: 16 October 2025

References

1. Balleine, B. W. & O'Doherty, J. P. Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychophamacology* **35**, 48–69 (2010).
2. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
3. Dickinson, A. Actions and habits: The development of behavioural autonomy. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* **308**, 67–78 (1985).
4. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
5. Collins, A. G. E. & Cockburn, J. Beyond dichotomies in reinforcement learning. *Nat. Rev. Neurosci.* **21**, 576–586 (2020).
6. Murakoshi, K. & Noguchi, T. Simulation of rat behavior by a reinforcement learning algorithm in consideration of appearance probabilities of reinforcement signals. *Biosystems* **80**, 83–90 (2005).
7. Boureau, Y.-L., Sokol-Hessner, P. & Daw, N. D. Deciding how to decide: Self-control and meta-decision making. *Trends Cogn. Sci.* **19**, 700–710 (2015).

8. Daw, N. D. Are we of two minds?. *Nat. Neurosci.* **21**, 1497–1499 (2018).
9. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
10. Afshar, N. M. et al. Reward-mediated, model-free reinforcement-learning mechanisms in Pavlovian and instrumental tasks are related. *J. Neurosci.* **43**, 458–471 (2023).
11. Akam, T. et al. The anterior cingulate cortex predicts future states to mediate model-based action selection. *Neuron* **109**, 149–163.e7 (2021).
12. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
13. Groman, S. M. et al. Neurochemical and behavioral dissections of decision-making in a rodent multistage task. *J. Neurosci.* **39**, 295–306 (2019).
14. Miller, K. J., Botvinick, M. M. & Brody, C. D. Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* **20**, 1269–1276 (2017).
15. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* **13**, e1005768 (2017).
16. Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Reversal learning and dopamine: a Bayesian perspective. *J. Neurosci.* **35**, 2407–2416 (2015).
17. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367 (2006).
18. Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
19. Hasz, B. M. & Redish, A. D. Deliberation and procedural automation on a two-step task for rats. *Front. Integr. Neurosci.* **12**, 30 (2018).
20. Decker, J. H., Otto, A. R., Daw, N. & Hartley, C. A. From creatures of habit to goal-directed learners. *Psychol. Sci.* **27**, 848–858 (2016).
21. Feher da Silva, C. & Hare, T. A. A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability. *PLoS ONE* **13**, e0195328 (2018).
22. Feher da Silva, C., Lombardi, G., Edelson, M. G. & Hare, T. Rethinking model-based and model-free influences on mental effort and striatal prediction errors. *Nat. Hum. Behav.* **7**, 956–969 (2023).
23. Potter, T. C. S., Bryce, N. V. & Hartley, C. A. Cognitive components underpinning the development of model-based learning. *Dev. Cogn. Neurosci.* **25**, 272–280 (2017).
24. Akam, T., Costa, R. & Dayan, P. Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Comput. Biol.* **11**, e1004648 (2015).
25. Bolenz, F., Kool, W., Reiter, A. M. & Eppinger, B. Metacontrol of decision-making strategies in human aging. *eLife* **8**, e49154 (2019).
26. Decker, J. H., Lourenco, F. S., Doll, B. B. & Hartley, C. A. Experiential reward learning outweighs instruction prior to adulthood. *Cogn. Affect. Behav. Neurosci.* **15**, 310–320 (2015).
27. Kool, W., Cushman, F. A. & Gershman, S. J. When does model-based control pay off?. *PLoS Comput. Biol.* **12**, e1005090 (2016).
28. Kool, W., Gershman, S. J. & Cushman, F. A. Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychol. Sci.* **28**, 1321–1333 (2017).
29. Smid, C. R. et al. Neurocognitive basis of model-based decision making and its metacontrol in childhood. *Dev. Cogn. Neurosci.* **62**, 101269 (2023).
30. Smid, C. R., Kool, W., Hauser, T. U. & Steinbeis, N. Computational and behavioral markers of model-based decision making in childhood. *Dev. Sci.* **26**, e13295 (2023).
31. Cowan, N. The many faces of working memory and short-term storage. *Psychon. Bull. Rev.* **24**, 1158–1170 (2017).
32. Collins, A. G. E., Ciullo, B., Frank, M. J. & Badre, D. Working memory load strengthens reward prediction errors. *J. Neurosci.* **37**, 4332–4342 (2017).
33. Dasgupta, I. & Gershman, S. J. Memory as a computational resource. *Trends Cogn. Sci.* **25**, 240–251 (2021).
34. Rmus, M., McDougle, S. D. & Collins, A. G. The role of executive function in shaping reinforcement learning. *Curr. Opin. Behav. Sci.* **38**, 66–73 (2021).
35. Yoo, A. H. & Collins, A. G. E. How working memory and reinforcement learning are intertwined: a cognitive, neural, and computational perspective. *J. Cogn. Neurosci.* **34**, 551–568 (2022).
36. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl Acad. Sci. USA.* **110**, 20941–20946 (2013).
37. Zuo, Z., Yang, L.-Z., Wang, H. & Li, H. Working memory guides action valuation in model-based decision-making strategy. *J. Cogn. Neurosci.* **37**, 86–96 (2025).
38. Cattell, R. B. Theory of fluid and crystallized intelligence: a critical experiment. *J. Educ. Psychol.* **54**, 1–22 (1963).
39. Cattell, R. B. *Abilities, Their Structure, Growth, and Action* (Houghton Mifflin, 1971).
40. Apšvalka, D., Cross, E. S. & Ramsey, R. Fluid intelligence and working memory support dissociable aspects of learning by physical but not observational practice. *Cognition* **190**, 170–183 (2019).
41. Wang, T., Ren, X., Altmeyer, M. & Schweizer, K. An account of the relationship between fluid intelligence and complex learning in considering storage capacity and executive attention. *Intelligence* **41**, 537–545 (2013).
42. Williams, B. A. & Pearlberg, S. L. Learning of three-term contingencies correlates with Raven scores, but not with measures of cognitive processing. *Intelligence* **34**, 177–191 (2006).
43. Schad, D. J. et al. Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Front. Psychol.* **5**, 1450 (2014).
44. Zhang, Z. Monte Carlo based statistical power analysis for mediation models: methods and software. *Behav. Res. Methods* **46**, 1184–1198 (2014).
45. Kool, W., Gershman, S. J. & Cushman, F. A. Planning complexity registers as a cost in metacontrol. *J. Cogn. Neurosci.* **30**, 1391–1404 (2018).
46. Kool, W. & Botvinick, M. Mental labour. *Nat. Hum. Behav.* **2**, 899–908 (2018).
47. Momennejad, I., Otto, A. R., Daw, N. D. & Norman, K. A. Offline replay supports planning in human reinforcement learning. *eLife* **7**, e32548 (2018).
48. Moran, R., Keramati, M. & Dolan, R. J. Model based planners reflect on their model-free propensities. *PLoS Comput. Biol.* **17**, e1008552 (2021).
49. Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* **24**, 751–761 (2013).
50. Simon, D. A. & Daw, N. D. Neural correlates of forward planning in a spatial decision task in humans. *J. Neurosci.* **31**, 5526–5539 (2011).
51. Velázquez-Vargas, C. A., Daw, N. D. & Taylor, J. A. The role of training variability for model-based and model-free learning of an arbitrary visuomotor mapping. *PLoS Comput. Biol.* **20**, e1012471 (2024).
52. Wunderlich, K., Dayan, P. & Dolan, R. J. Mapping value based planning and extensively trained choice in the human brain. *Nat. Neurosci.* **15**, 786–791 (2012).
53. Kim, D., Park, G. Y., O'Doherty, J. P. & Lee, S. W. Task complexity interacts with state-space uncertainty in the arbitration between

model-based and model-free learning. *Nat. Commun.* **10**, 5738 (2019).
54. Otto, A. R., Skatova, A., Madlon-Kay, S. & Daw, N. D. Cognitive control predicts use of model-based reinforcement learning. *J. Cogn. Neurosci.* **27**, 319–333 (2015).
55. Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D. & Dolan, R. J. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron* **80**, 914–919 (2013).
56. Eckstein, M. K., Wilbrecht, L. & Collins, A. G. What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Curr. Opin. Behav. Sci.* **41**, 128–137 (2021).
57. Gershman, S. J. *Reinforcement Learning and Causal Models* (ed. Waldmann, M. R.) Vol. 1 (Oxford University Press, 2017).
58. Rummery, G. & Niranjan, M. *On-Line Q-Learning Using Connectionist Systems* (Cambridge University, 1994).
59. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 1998).
60. Marc, J. M. AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c) (R package version 2.3-1). https://cran.r-project.org/web/packages/AICcmodavg/index.html (2020).
61. Lüdecke, D. ggeffects: Tidy data frames of marginal effects from regression models. *J. Open Source Softw.* **3**, 772 (2018).
62. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2014).

Acknowledgements

This study was funded by the National Natural Science Foundation of China (Grant No. 31370020), the Natural Science Foundation of Liaoning Province of China (2022-KF-26-01). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

C.Y.: Conceptualization, data curation, formal analysis, investigation, methodology, writing—original draft preparation, writing—review & editing. T.L.: Conceptualization, methodology, investigation, project administration, funding acquisition, supervision, writing—review & editing. M.W.: Formal analysis, methodology, writing—review & editing. X.L.: Conceptualization, methodology, investigation, supervision, writing—review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41539-025-00363-w.

Correspondence and requests for materials should be addressed to Tongran Liu.