Title: The neurocognitive role of working memory load when Pavlovian motivational control affects instrumental learning

Authors:

Heesun Park[1], Hoyoung Doh[1], Harhim Park[1], Woo-Young Ahn[1,2]

[1]Department of Psychology, Seoul National University, Seoul, Korea 08826

[2]Department of Brain and Cognitive Sciences, Seoul National University, Seoul, Korea 08826

Corresponding author:

Woo-Young Ahn, Ph.D.
Department of Psychology
Seoul National University
Seoul, Korea 08826
Tel: +82-2-880-2538, Fax: +82-2-877-6428. E-mail: wahn55@snu.ac.kr

Number of pages: 42

Number of Figures & Tables: 5/1

Number of words in the abstract: 272

Number of words in the introduction: 1018

Number of words in the discussion: 2223

Conflict of Interest: The authors declare no competing financial interests.

Abstract

Humans and animals learn optimal behaviors by interacting with the environment. Research suggests that a fast, capacity-limited working memory (WM) system and a slow, incremental reinforcement learning (RL) system jointly contribute to instrumental learning. Situations that strain WM resources alter several decision-making processes and the balance between multiple decision-making systems: under WM loads, learning becomes slow and incremental, while reward prediction error (RPE) signals become stronger; the reliance on computationally efficient learning increases as WM demands are balanced against computationally costly strategies; and action selection becomes more random. Meanwhile, instrumental learning is known to interact with Pavlovian learning, a hard-wired system that motivates approach to reward and avoidance of punishment. However, the neurocognitive role of WM load during instrumental learning under Pavlovian influence remains unknown, while conflict between the two systems sometimes leads to suboptimal behavior. Thus, we conducted a functional magnetic resonance imaging (fMRI) study (N = 49) in which participants completed an instrumental learning task with Pavlovian–instrumental conflict (the orthogonalized go/no-go task); WM load was manipulated with dual-task conditions. Behavioral and computational modeling analyses revealed that WM load compromised learning by reducing the learning rate and increasing random choice, without affecting Pavlovian bias. Model-based fMRI analysis revealed that WM load strengthened RPE signaling in the striatum. Moreover, under WM load, the striatum showed weakened connectivity with the ventromedial and dorsolateral prefrontal cortex when computing reward expectations. These results suggest that the limitation of cognitive resources by WM load decelerates instrumental learning through the weakened cooperation between WM and RL; such limitation also makes action selection more random, but it does not directly affect the balance between instrumental and Pavlovian systems.

Introduction

The process of learning about the environment from experience and making adaptive decisions involves multiple neurocognitive systems, among which reinforcement learning (RL) and working memory (WM) systems are known to significantly contribute to learning (Collins & Frank, 2012; Huys et al., 2021; Rmus et al., 2021). RL processes facilitate "incremental" learning from the discrepancy between actual and predicted rewards, known as reward prediction error (RPE); RL is regarded as a slow but steady process (Sutton & Barto, 2018). Dopaminergic activity in the basal ganglia conveys RPEs (Bornstein & Daw, 2011; Khamassi et al., 2005; Montague et al., 1996; Niv, 2009; Schultz, 1997, 1998; Schultz et al., 1997), and human imaging studies have found that blood-oxygen-level-dependent (BOLD) signals in the striatum are correlated with RPEs (Garrison et al., 2013; J. O'Doherty et al., 2004; J. P. O'Doherty et al., 2003).

In addition to RL, WM is a crucial component in learning. In particular, WM allows the rapid learning of actions via retention of recent stimulus-action-outcome associations, while RL constitutes a slow learning process (Collins, Ciullo, et al., 2017; Collins, 2018; Collins & Frank, 2012; Yoo & Collins, 2022). WM can also offer various inputs to RL, such as reward expectations (Collins & Frank, 2018) and models of the environment (Dayan, 2009; Dolan & Dayan, 2013; Tanaka et al., 2008; Valentin et al., 2007) as well as complex states and actions (Collins & Shenhav, 2021; Rmus et al., 2021). In the brain, the WM system is presumably associated with sustained neural activity throughout the dorsolateral prefrontal cortex (dlPFC) and prefrontal cortex (PFC) (Baddeley & Hitch, 1974; Barbey et al., 2013; Curtis & D'Esposito, 2003; Funahashi, 2006; Funahashi & Kubota, 1994; Rottschy et al., 2012).

Because RL and WM cooperate to promote successful learning, the deterioration of either system can alter the learning and balance between the two systems. In particular, increasing WM load during learning and decision-making can lead to various consequences through the depletion of WM resources. For example, first, instrumental learning becomes slow and incremental under WM load (Collins, 2018; Collins, Albrecht, et al., 2017; Collins &

Frank, 2012; McDougle & Collins, 2020). Limited resources in the WM system cause WM contribution to decline while the RL contribution increases, causing learning to occur more slowly and strengthening the RPE signal in the brain (Collins, Ciullo, et al., 2017; Collins & Frank, 2018). Second, among the multiple RL systems that use varying degrees of WM resources, the demands of WM can be balanced against computationally costly strategies. Otto et al. demonstrated that under WM load, the reliance on computationally efficient model-free learning was increased, compared with model-based learning (Otto et al., 2013). Lastly, limited WM resources may cause action selection to become more random and inconsistent. Different values must be compared to inform decision-making during the action selection stage (Rangel et al., 2008), but several studies have reported that WM load may interrupt these processes without affecting valuation itself (Franco-Watkins et al., 2006, 2010; Olschewski et al., 2018).

While reductions of WM resources substantially alter instrumental learning, another important factor known to shape instrumental learning is the Pavlovian system. Through the motivation of hard-wired responses, such as active responses to appetitive cues and inhibitory responses toward aversive cues (Dickinson & Balleine, 2002; Mackintosh, 1983; Wasserman et al., 1974; Wasserman & Miller, 1997), the Pavlovian system may facilitate certain instrumental behaviors and impede others. This bias in instrumental learning is known as Pavlovian bias (Breland & Breland, 1961; Dayan et al., 2006; Hershberger, 1986; Williams & Williams, 1969). Pavlovian bias is generally presumed to be associated with maladaptive behaviors such as substance use disorder and compulsivity-related disorders (Everitt & Robbins, 2005; Garbusow et al., 2014, 2016; Glasner et al., 2005; Lüscher et al., 2020).

Although it is well known that the enhancement of WM load alters instrumental learning in several ways, it remains unclear how WM load changes instrumental learning when it is under Pavlovian influence. To investigate this relationship, we conducted a functional magnetic resonance imaging (fMRI) study in which participants completed an instrumental learning task that involved Pavlovian–instrumental conflicts (Guitart-Masip et al., 2012), with and without additional WM load.

We tested the following three hypotheses. First, if the role of WM in learning is unaffected by Pavlovian influence, WM load will lead to slower learning and increased striatal RPE signals, consistent with previous findings (Collins, 2018; Collins, Ciullo, et al., 2017; Collins & Frank, 2012, 2018). Second, if WM load leads to a computational trade-off between Pavlovian and instrumental learning, as model-free and model-based learning (Otto et al., 2013), WM load will enhance Pavlovian bias because the Pavlovian system is known to require fewer resources and to be computationally efficient as an evolutionarily embedded system that learns values as a function of cues, regardless of actions (Dayan et al., 2006). We also presumed that neural signaling associated with Pavlovian bias would increase under WM load. We focused on regions of the basal ganglia, such as the striatum and substantia nigra/ventral tegmental area (SN/VTA), which are considered important in Pavlovian bias (Boer et al., 2018; Chowdhury et al., 2013; Frank et al., 2004; Guitart-Masip et al., 2012; Guitart-Masip, Duzel, et al., 2014). Third, if the contribution of WM to consistent action selection remains consistent, WM load will cause action selection to become more random, as in previous studies (Franco-Watkins et al., 2006, 2010; Olschewski et al., 2018). We tested whether the value comparison signal in the brain would decrease under WM load because consistent action selection may be associated with the extent to which value difference information is utilized during the decision-making process (Gläscher & O'Doherty, 2010; Rangel et al., 2008).

Our behavioral and computational modeling results revealed that Pavlovian bias did not increase under WM load, while learning decelerated and action selection became increasingly random; these findings supported hypotheses 1 and 3 but not 2. Increased striatal RPE signaling suggests that the increased contribution of RL and decreased contribution of WM may explain slower learning. Further analyses revealed weakened connectivity between the striatal and prefrontal regions under WM load, suggesting diminished cooperation between the WM and RL systems.

Results

The participants (N = 56) underwent fMRI imaging while performing an instrumental learning task under a control condition and a WM load condition (**Figure 1**). In the control condition, they participated in the orthogonalized go/no-go (GNG) task (Guitart-Masip et al., 2012), a model-free learning task that contained Pavlovian–instrumental conflicts. In the WM load condition, a 2-back task was added to the GNG task; the modified task was named the working memory go/no-go (WMGNG) task (see Materials and Methods for more detail).

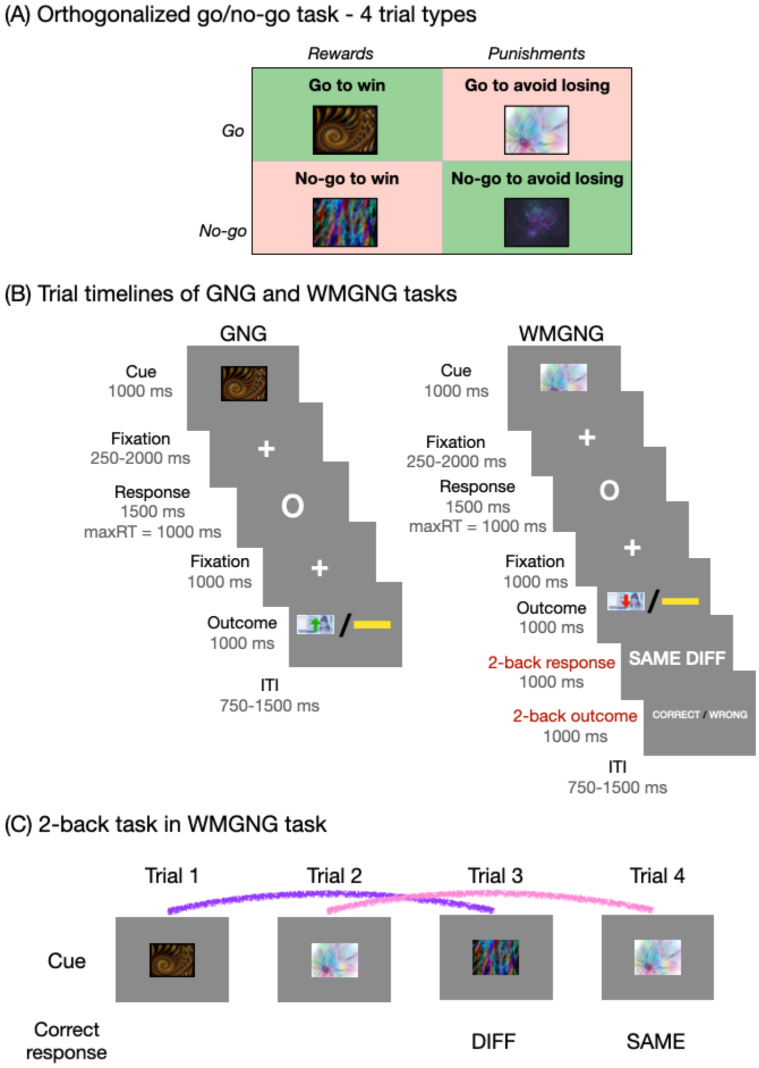

Figure 1. The GNG and WMGNG tasks. (A) In both tasks, four fractal cues indicated the combination of action (go/no-go) and valence at the outcome (win/loss). (B) In each trial, a fractal cue was presented, followed by a variable delay. After the delay, actions were required in response to a circle, and

participants had to decide whether to press a button. After an additional brief delay, the probabilistic outcome was presented, indicating monetary reward (green upward arrow on a ₩1000 bill) or monetary punishment (red downward arrow on a ₩1000 bill). A yellow horizontal bar indicated no win or loss. In the WMGNG task, the original GNG task was followed by a 2-back response and 2-back outcome phases. (C) The participants were asked to indicate whether the cue in the current trial was identical to the cue in the two preceding trials. Here, because the cue in trial 3 differed from the cue in trial 1, "DIFF" was the correct response. Similarly, because the cue in trial 4 was identical to the cue in trial 3, "SAME" was the correct response. The lines mark two cues for comparison: the purple line indicates that the cues differ, while the pink line indicates that the cues are identical.

Task performance: Decreased performance and learning speed under WM load

Comparison of overall task accuracy between the two tasks confirmed that our dual-task manipulation with a 2-back task successfully imposed WM load. Participants performed better in the GNG task (*M*=0.80, *SD*=0.12) than in the WMGNG task (*M*=0.72, *SD*=0.16), as illustrated in **Figure 2A** (paired t-test, *t*(48)=3.86, *p*<0.001, *d*=0.55). We also confirmed that participants exhibited go bias and Pavlovian bias in both tasks, thus replicating the findings of earlier studies (Adams et al., 2020; Betts et al., 2020; Boer et al., 2018; Ereira et al., 2021; Guitart-Masip, Economides, et al., 2014; Guitart-Masip et al., 2012; Perosa et al., 2020; Richter et al., 2014, 2021). Two-way ANOVA on accuracy, with the factors action (go/no-go) and valence (reward/punishment) as repeated measures for both tasks, revealed a main effect of action (*F*(48)=6.05, *p*=0.018, η^2=0.03 in GNG task, *F*(48)=9.44, *p*=0.003, η^2=0.04 in WMGNG task) and action by valence interaction (*F*(48)=22.43, *p*<0.001, η^2=0.12 in the GNG task, *F*(48)=30.59, *p*<0.001, η^2=0.10 in the WMGNG task); it showed no effect of valence (*F*(48)=0.00, *p*=0.99, η^2=0.00 in the GNG task, *F*(48)=2.77, *p*=0.103, η^2=0.01 in the WMGNG task). In both tasks (**Figure 2B**), participants exhibited superior performances in "go to win" and "no-go to avoid losing" conditions (i.e., Pavlovian-congruent conditions; blue columns) than in "no-go to win" and "go to avoid losing" trials (i.e., Pavlovian-incongruent conditions; red columns). Specifically, in the GNG task, accuracy was higher in the "go to win" (*M*=0.92, *SD*=0.12) than "no-go to win" condition (*M*=0.69, *SD*=0.35) (paired t-test, *t*(48)=4.13, *p*<0.001, *d*=0.59), and in the “no-go to avoid losing” (*M*=0.85, *SD*=0.13) than in the “go to avoid losing” condition (*M*=0.76, *SD*=0.18) (paired t-test, *t*(48)=3.29, *p*=0.002, *d*=0.47). Similarly, in the WMGNG task, accuracy was higher in the “go to win” (*M*=0.82, *SD*=0.25) than in the “no-go to win” condition (*M*=0.57, *SD*=0.34) (paired t-test, *t*(48)=4.82, *p*<0.001, *d*=0.69), and in the “no-go to avoid losing” (*M*=0.79, *SD*=0.16) than in the “go to avoid losing” condition (*M*=0.72, *SD*=0.19) (paired t-test, *t*(48)=2.51, *p*=0.015, *d*=0.36).

Next, we tested the effect of WM load on learning speed (hypothesis 1, **Figure 2C**). While the learning curves indicated that participants learned during both tasks, the learning curve was slower in the WMGNG task than in the GNG task (i.e., WM load reduced learning speed and overall accuracy). To test the effect of WM load on Pavlovian bias (hypothesis 2, **Figure 2D**), we quantified Pavlovian bias by subtracting the accuracy in Pavlovian-incongruent conditions (“no-go to win” and “go to avoid losing”) from accuracy in Pavlovian-congruent conditions (“go to win” and “no-go to avoid losing”), then compared it between the two tasks. No significant difference in Pavlovian bias was observed between the GNG and WMGNG tasks, confirming that WM load did not affect Pavlovian bias.

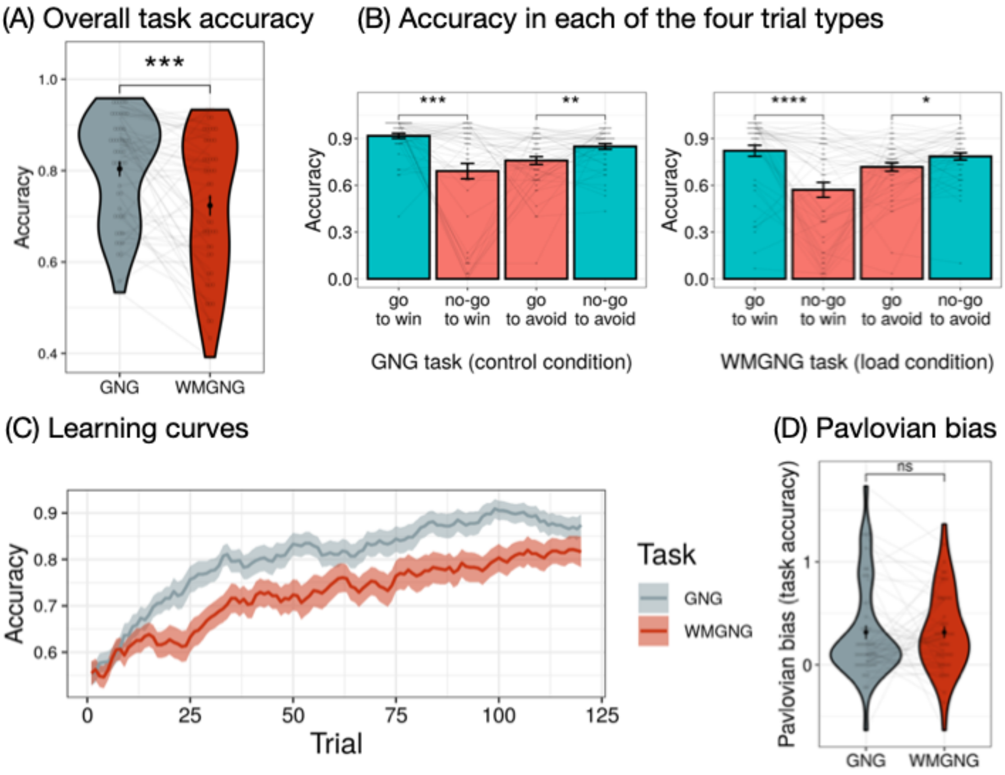

Figure 2. Task performance. (A) Task accuracies (mean percentages of correct responses) in the GNG and WMGNG tasks show that participants performed better in the GNG task than in the WMGNG task. (B) Accuracy in each of the four trial types between the two tasks demonstrated that participants performed better in “go to win” and “no-go to avoid losing” trials (Pavlovian-congruent, blue) than in “no-go to win” and “go to avoid losing” trials (Pavlovian-incongruent, red). (C) The learning curve (i.e., the increase in accuracy across trials) was slower in the WMGNG task than in the GNG task. Note that moving average smoothing was applied with filter size 5 to remove the fine variation between time steps. Lines indicate group means and ribbons indicate ± standard errors of the mean. (D) Pavlovian bias was calculated by subtracting accuracy in Pavlovian-incongruent conditions (“no-go to win” + “go to avoid losing”) from accuracy in Pavlovian-congruent conditions (“go to win” + “no-go to avoid losing”). No significant difference in Pavlovian bias was observed between the GNG and WMGNG tasks. (A)-(B), (D) Dots indicate group means and error bars indicated ± standard errors of the mean. Gray dots indicate individual accuracies; lines connect a single participant’s performances. Asterisks indicate the results of pairwise t-tests. **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Computational modeling: WM load influences learning rate and irreducible noise

We used a computational modeling approach to test the three hypotheses. For this purpose, we developed eight nested models that assumed different learning rate, Pavlovian bias, or irreducible noise parameters under WM load. These models were fitted to the data using hierarchical Bayesian analysis, then compared using the leave-one-out information criterion (LOOIC), where a lower LOOIC value indicated better out-of-sample predictive accuracy (i.e., better fit) (Vehtari et al., 2017). Importantly, the use of computational modeling allowed us to test hypothesis 3 regarding whether WM load would increase random choices; this would have not been possible if we had performed behavioral analysis alone.

Based on earlier studies (Cavanagh et al., 2013; Guitart-Masip et al., 2012), we constructed a baseline model (model 1) that used a Rescorla-Wagner updating rule and contained learning rate (ε), Pavlovian bias, irreducible noise, go bias, and separate parameters for sensitivity to rewards and punishments (Materials and Methods). In the model, state-action values are updated with the prediction error; learning rate (ε) modulates the impact of the prediction error. Reward/punishment sensitivity (ρ) scales the effective size of outcome values. Go bias (b) and cue values weighted by Pavlovian bias (π) are added to the value of go choices. Here, as the Pavlovian bias parameter increases, the go tendency increases under the reward condition whereas the go tendency is reduced under the punishment condition; this results in an increased no-go tendency. Computed action weights are used to estimate action probabilities, and irreducible noise (ξ) determines the extent to which information about action weights is utilized to make decisions. As irreducible noise increases, action probabilities will be less reflective of action weights, indicating that action selection will become increasingly random.

In models 2, 3, and 4, we assumed that WM load affects only one parameter. For example, in model 2, a separate Pavlovian bias parameter (π_{wm}) was assumed for the WM load condition. Models 3 and 4 assumed different learning rates (ε_{wm}) and irreducible noise (ξ_{wm}) parameters in their respective WM load conditions. In models 5, 6, and 7, we assumed that WM load would affect two parameters: model 5 had different Pavlovian bias (π_{wm}) and learning rate (ε_{wm}); model 6 had different Pavlovian bias (π_{wm}) and irreducible noise (ξ_{wm}); and model 7 had different learning rate (ε_{wm}) and irreducible noise (ξ_{wm}). Finally, model 8 was the full model, in which all three parameters were assumed to be affected by WM load.

The full model (model 8) was the best model (**Figure 3A**). In other words, it demonstrated that participant behavior could be best explained when separate parameters were included for Pavlovian bias, learning rate, and irreducible noise parameters. Next, we analyzed the parameter estimates of the best-fitting model; we focused on comparing the posterior distributions of the parameters that were separately fitted in the two tasks (**Figure 3B**). The parameters were considered credibly different from each other if the 95% highest density intervals (HDI) of the two distributions showed no overlap (Kruschke, 2014). **Figure 3B** illustrates that Pavlovian bias was not credibly different between the two tasks, consistent with the lack of support for hypothesis 2 (Pavlovian bias) in the behavioral results. Conversely, the learning rate was credibly lower, while irreducible noise was credibly greater in the WMGNG than in the GNG task. These results support hypothesis 1 (i.e., WM load will reduce learning rate) and hypothesis 3 (i.e., WM load will increase random choices). While the best model was the full model that assumed separate Pavlovian bias in the two tasks, no credible group difference was observed between these parameters. This is presumably because the full model was able to capture individual variations among participants (**Figure S3**), despite the lack of credible difference in the group-level estimates between the two tasks. As expected, the 95% HDIs of go bias, reward sensitivity, and punishment sensitivity did not include zero, indicating that the participants exhibited go bias and reward/punishment sensitivity (see Supplementary Material for the posterior distributions of individual parameters; **Figure S2-S5**).

To further compare choice randomness between the two tasks, we examined the extent to which choices were dependent on value discrepancies between the two options. We first plotted the percentage of go choices for the GNG and WMGNG tasks by varying the quantiles of differences in action weight between the "go" and "no-go" actions (W_{go} - W_{nogo}) (**Figure 4A**). The trial-by-trial action weights were extracted from the best-fitting model. Higher quantiles corresponded to a greater "go" action weight than "no-go" action weight. Overall, the go ratio increased from the first to the tenth quantile, indicating that the value differences between the "go" and "no-go" actions affected participants' choices. This result further illustrates the difference between the two tasks: the increase in the go ratio was steeper in the GNG task than in the WMGNG task. In particular, the go ratio significantly differed between the two tasks for the first (*t*(48)=-3.59, *p*=0.001, *d*=0.51), second (*t*(48)=-3.23, *p*=0.002, *d*=0.46), third (*t*(48)=-2.55, *p*=0.014, *d*=0.36), eighth (*t*(48)=2.95, *p*=0.005, *d*=0.42), and tenth (*t*(48)=2.76, *p*=0.008, *d*=0.39) quantiles. Thus, under WM load, participants were less sensitive to the significant value difference between "go" and "no-go".

To compare these patterns in a different way and further explore the extent to which performance was dependent on choice difficulty, we plotted accuracies for the two tasks and for different quantiles of the absolute value differences ($|W_{go} - W_{nogo}|$; **Figure 4B**). We assumed that the choices would become easier when the absolute value difference was increased because a small value difference makes it difficult to choose between options. Overall, the accuracy increased from the first to the tenth quantile, indicating that participants performed better as the choices became easier. This result further illustrates the difference between the two tasks: the increase in accuracy was steeper in the GNG task than in the WMGNG task. Specifically, the accuracy significantly differed between the two tasks for the fifth ($t(48)=4.12$, $p<0.001$, $d=0.59$), sixth ($t(48)=2.95$, $p=0.005$, $d=0.42$), seventh ($t(48)=2.44$, $p=0.018$, $d=0.35$), eighth ($t(48)=3.13$, $p=0.003$, $d=0.45$), ninth ($t(48)=2.87$, $p=0.006$, $d=0.41$), and tenth ($t(48)=2.55$, $p=0.014$, $d=0.36$) quantiles. Thus, participants performed worse in the WM load condition than in the control condition when choices were easier. Overall, **Figure 4** demonstrates that WM load reduced the effect of the value difference on participants, indicating increased choice randomness.

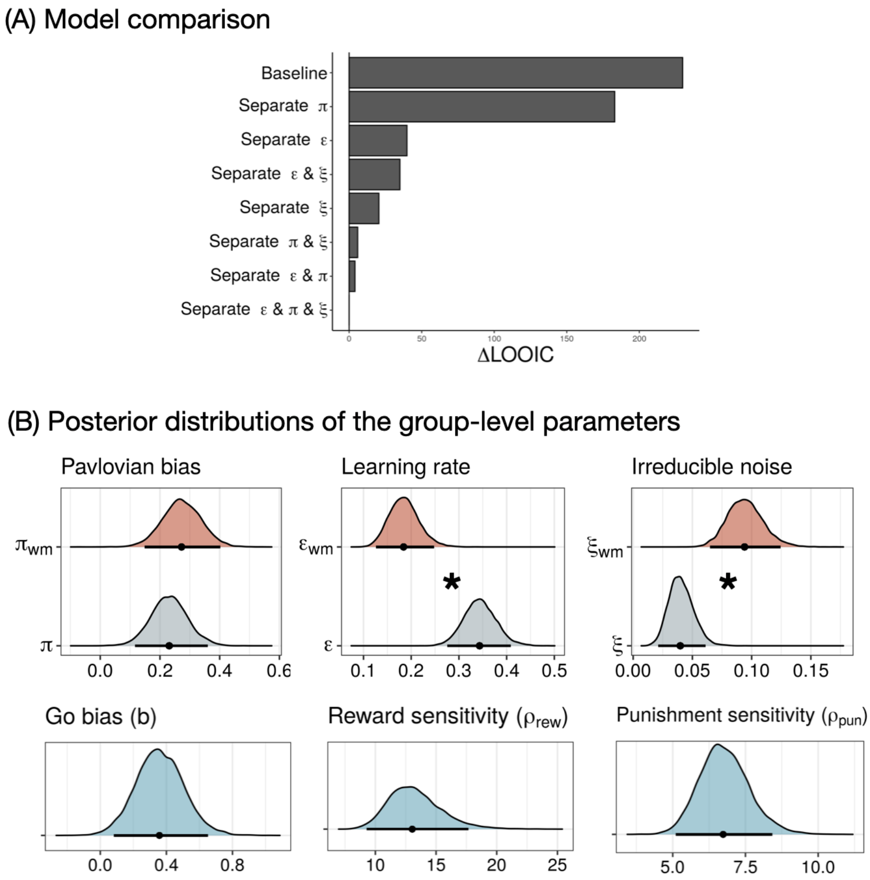

Figure 3. Model comparison results and posterior distribution of the group-level parameters of the best-fitting model. (A) Relative LOOIC difference indicates the difference in LOOIC between the best-fitting model and each of the other models. The best-fitting model was the full model, which assumed separate Pavlovian bias, learning rate, and irreducible noise in GNG and WMGNG tasks. Lower LOOIC indicates better model fit. (B) Posterior distributions of group-level parameters from the best-fitting model.

Learning rate and irreducible noise estimates were credibly different in the GNG and WMGNG tasks, while Pavlovian bias estimates were not. Dots indicate medians and bars indicate 95% HDIs. Asterisks indicate that the 95% HDIs of the two parameters' posterior distributions do not overlap (i.e., differences are credible).

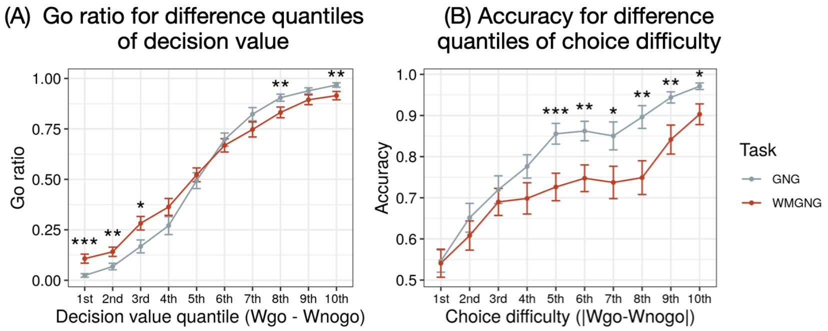

Figure 4. Choice consistency. (A) Mean percentage of go choices for different quantiles of action weight differences (W_{go} - W_{nogo}) between "go" and "no-go" choices, where higher quantiles indicate higher decision values for "go" choices. Under WM load, the increase in go ratio according to quantile was less steep. (B) Mean accuracies for different quantiles of absolute value differences ($|W_{go}$ - $W_{nogo}|$), where higher quantiles indicate larger value differences between two options or easier choices. Under WM load, the increase in accuracy according to quantile was less steep. (A)-(B) Dots are group means, and error bars are ± standard errors of the mean. Asterisks show the results of pairwise t-tests. **** *p* < *0.0001, *** p < 0.001, ** p < 0.01, * p < 0.05.*

Larger RPE signals in the striatum and weakened connectivity with prefrontal regions under WM load

Behavioral analysis revealed that WM load caused learning to occur more slowly but did not affect Pavlovian bias. The computational approach confirmed that the learning rate decreased; Pavlovian bias did not change under the load; and WM load led to increased choice randomness. Here, we sought to investigate the underlying neural correlates of these effects of WM load on learning rate, Pavlovian bias, and random action selection. First, we hypothesized that RPE signaling in the striatum would increase under WM load (Collins, Ciullo, et al., 2017; Collins & Frank, 2018). We conducted a model-based fMRI analysis using RPE as a regressor derived from the best-fitting model (see Materials and Methods for the full general linear models (GLMs) and regressor specifications). The RPE signal in the striatal region of interest (ROI) was significantly greater in the WMGNG task than in the GNG task (contrast: RPE in WMGNG > RPE in GNG, MNI space coordinates *x* = 13, *y* = 14, *z* = -3, *Z* = 3.96, $p < 0.05$ small-volume corrected (SVC), **Figure 5A**, **Table S4**). This supports hypothesis 1, which predicts an increased contribution of the RL system and decreased contribution of the WM system, to learning under WM load. We also tested hypothesis 2 regarding Pavlovian bias, but we found no main effect of Pavlovian bias between the GNG and WMGNG tasks (WMGNG > GNG [Pavlovian-congruent > Pavlovian-incongruent]) within the striatum or SN/VTA ($p < 0.05$ SVC). Note that previous studies showed no significant result for the same contrast (Pavlovian-congruent > Pavlovian-incongruent) within the same regions (Guitart-Masip et al., 2012). With regard to hypothesis 3 concerning random choices, we observed no main effect of WM load on random choice (WMGNG > GNG [W_{chosen} - $W_{unchosen}$]) within the ventromedial prefrontal cortex (vmPFC; $p < 0.05$ SVC). See Supplementary Material for further details regarding these findings (**Table S5**).

Increased RPE signals under WM load may indicate reduced WM contribution and increased RL contribution to learning because of the load, suggesting diminished cooperation between the two systems for learning. Therefore, we conducted a psychophysiological interaction (PPI) analysis (Friston et al., 1997) using the gPPI toolbox (McLaren et al., 2012) to test whether functional connectivity between areas associated with RL and WM systems would weaken under WM load. Specifically, we explored differences between the two tasks in terms of functional coupling between the striatum, which showed increased RPE signaling under WM load, and other regions when computing reward expectations. The striatum showed significantly decreased connectivity with the vmPFC (MNI space coordinates *x* = 13, *y* = 56, *z* = 0, *Z* = −4.90, $p < 0.05$ whole-brain cluster-level familywise error rate (FWE)) and dlPFC (MNI space coordinates *x* = -20, *y* = 63, *z* = 23, *Z* = −4.24, $p < 0.05$ whole-brain cluster-level FWE, **Figure 5B**, **Table S6**) in the WMGNG task, compared with the GNG task.

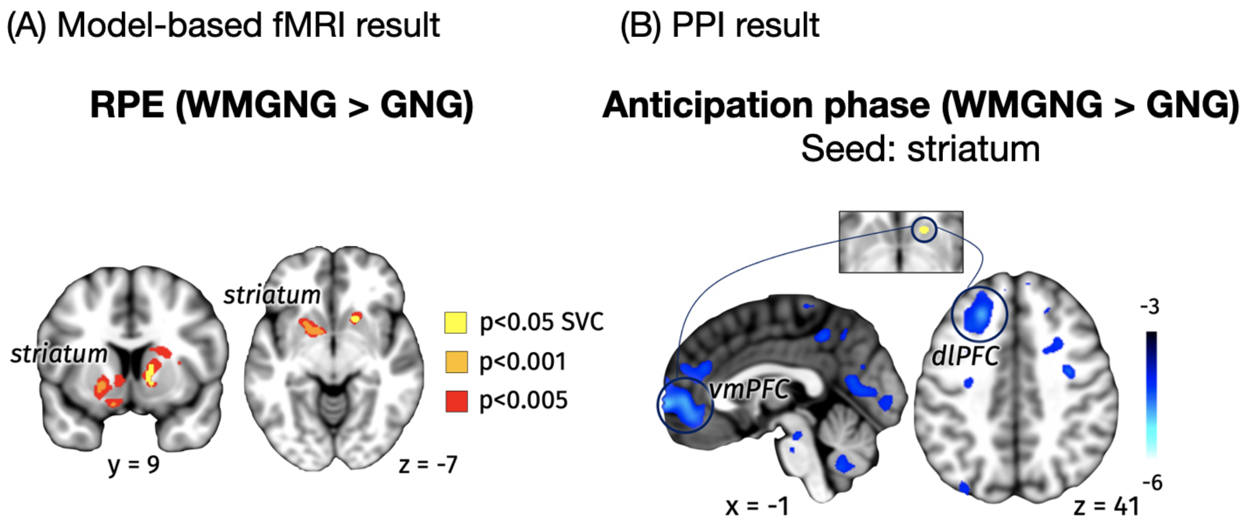

Figure 5. fMRI results. (A) RPE signaling in the striatum was stronger in the WMGNG task than in the GNG task. Effects that were significant at p < 0.05 (SVC) are shown in yellow. (B) Functional connectivity between the striatum (seed region, top) and prefrontal regions, including vmPFC (bottom left) and dlPFC (bottom right), was weaker in the WMGNG task than in the GNG task when computing reward expectation (p < 0.05, whole-based cluster-level FWE). Overlays are shown with a threshold of p < 0.001 (uncorrected). Color scale indicates t-values.

Discussion

In this study, our main objective was to elucidate the neurocognitive effects of WM load on instrumental learning that involves Pavlovian–instrumental conflicts. We hypothesized that under WM load, 1) learning rate would decrease and RPE signals would become stronger, 2) Pavlovian bias would increase, and 3) action selection would become increasingly random. First, we found that the limitation of WM resources according to WM load led to a decrease in the learning rate and increases in striatal RPE signals. The striatum, which subsequently showed stronger RPE signals under WM load, demonstrated weakened functional connectivity with prefrontal regions including the dlPFC and vmPFC, during reward prediction. WM load also increased random action selection. However, Pavlovian bias did not increase under WM load, suggesting that WM load did not affect the balance between Pavlovian and instrumental systems.

Decreased contribution of the WM system and increased contribution of the RL system under WM load

The effect of WM load on instrumental learning remained consistent despite Pavlovian bias. In particular, our behavioral analysis revealed a deceleration in learning speed under WM load (**Figure 2C**); modeling analysis confirmed that WM load reduced learning rate (**Figure 3**).

These findings can be attributed to the reduced contribution of the WM system and increased contribution of the striatal RL system, consistent with previous findings that WM improves learning efficiency (in parallel with RL), as well as reward prediction precision in RL processes. First, as a rapid and immediate learning system, WM learns in parallel with the slow and incremental RL system by directly storing associations between states and actions (Collins, 2018; Collins, Ciullo, et al., 2017; Collins & Frank, 2012; Tsujimoto & Sawaguchi, 2004; Yoo & Collins, 2022). Specifically, WM and RL systems compete with each other based on their reliability in a given situation. Under WM load, the fast and capacity-limited WM system becomes less reliable; thus, the slow and incremental RL system supersedes the WM system, causing learning to occur more slowly and incrementally (Collins, 2018; Collins, Albrecht, et al., 2017; Collins & Frank, 2012). Second, RL computations themselves are intertwined with WM; WM can represent feed reward expectations to the RL system (Ballard et al., 2011; Kahnt et al., 2011; D. Lee & Seo, 2007; Wallis & Miller, 2003) and improve reward prediction precision, which can reduce RPE and improve learning efficiency (Collins, Ciullo, et al., 2017; Collins & Frank, 2018). In our study, WM load, which limited the contribution of WM, may have increased the striatal RL contribution while reducing the accuracy of RL reward computation. Consistent with this interpretation, we found that RPE signaling in the striatum – a marker of RL – was strengthened under WM load (**Figure 5A**). This is consistent with previous findings that RPE-associated neural signals were increased under higher WM load (Collins, Ciullo, et al., 2017; Collins & Frank, 2018).

Overall, these findings suggest that WM load led to reduced cooperation between RL and WM by interrupting and reducing the contribution of WM. This notion is further supported by the finding that the striatum showed weakened functional connectivity with the dlPFC during reward prediction under WM load (**Figure 5A**). Taken together, these findings suggest that WM load may have weakened the interplay between WM in the dlPFC and RL in the striatum during the value estimation process, which subsequently led to stronger RPE signals. However, further research is necessary to demonstrate the directionality of functional connectivity between the two systems during reward prediction; frontostriatal connectivity is reportedly bidirectional, such that the striatum may also provide prefrontal regions with inputs that relate to reward information (Park et al., 2010; Pasupathy & Miller, 2005).

Notably, we observed weakened connectivity between the vmPFC and the striatum. The vmPFC has been identified as a critical neural correlate of value-based decision-making; it integrates reward predictions (Kahnt et al., 2011), represents value signals or decision value (Economides et al., 2014; Lim et al., 2011; O'Doherty, 2011; Smith et al., 2010), and affects reward anticipation/processing in the striatum (Hiser & Koenigs, 2018; Pujara et al., 2016). Our findings suggest that value integration through the cortico-striatal loop was also weakened under WM load.

No effect of WM load on Pavlovian bias

Contrary to our hypothesis, WM load did not influence Pavlovian bias. Behavioral and modeling results showed that Pavlovian bias did not significantly differ between the GNG and WMGNG tasks (**Figure 2D; Figure 3**), while fMRI analysis revealed that neural signaling associated with Pavlovian bias did not significantly differ between the two tasks (**Table S6**). These findings indicate that the brain did not exhibit greater reliance on the computationally efficient system under WM load, in contrast to the results of previous studies (Otto et al., 2013). We identified two possible explanations for this discrepancy. First, instrumental and Pavlovian learning require similar amounts of WM resources; second, the WM system may not be involved in modulating the balance between Pavlovian and instrumental systems.

In the first potential explanation, the amounts of WM resources may be similar for Pavlovian and instrumental learning (especially model-free learning), in contrast to model-based and model-free learning. Model-based learning system constructs an internal model to compute the values of actions; thus, it requires greater WM resources to compute and retain the model online (Balleine & O'doherty, 2010; Daw et al., 2005, 2011; Dolan & Dayan, 2013; Keramati et al., 2011). However, the model-free system simply uses the action-reward association history to compute action values (i.e., "cached values"), and does not require the internal model (Balleine & O'doherty, 2010; Daw et al., 2005; Dickinson, 1985). Pavlovian learning is similar to model-free learning but differs in terms of the dimensions for value learning–the Pavlovian system learns state-outcome associations, while the instrumental system learns state-action-outcome associations (Dayan et al., 2006; Dorfman & Gershman, 2019). Therefore, the difference in WM demands between model-based and model-free learning could be significantly greater than the difference between model-free instrumental and Pavlovian learning. In our task, in particular, the instrumental learning was model-free; both instrumental and Pavlovian systems were required to learn the associations without prior information. Thus, the difference in WM demands may not have been sufficient to trigger a trade-off between the two learning systems. Rather than depending more on Pavlovian learning which has little computational benefit in our task, the participants may simply have compromised overall learning.

In the second potential explanation, WM resources may be unimportant with respect to modulating the Pavlovian–instrumental interaction, despite earlier studies' suggestions to the contrary. Several studies have proposed that prefrontal WM control systems are crucial for controlling Pavlovian bias. Electroencephalogram studies demonstrated that midfrontal theta oscillations are important for controlling Pavlovian bias (Cavanagh et al., 2013; Swart et al., 2018), suggesting top-down prefrontal control over Pavlovian bias (Cavanagh et al., 2013). Furthermore, recruitment of the inferior frontal gyrus (IFG) is involved in appropriate response inhibition, helping to overcome Pavlovian bias (Guitart-Masip et al., 2012). Finally, there is indirect evidence that administration of levodopa, which increases dopamine levels, reduced

Pavlovian influences on instrumental learning; such a reduction was speculated to result from increased dopamine levels in the PFC, which may have facilitated the operation of prefrontal WM functions (Guitart-Masip, Economides, et al., 2014). A related finding suggested that genetic determinants of prefrontal dopamine function may be important in overcoming Pavlovian bias (Richter et al., 2021).

While the results of the present study appear to contradict these findings, several complex possibilities exist. In particular, although previous findings implied the involvement of prefrontal mechanisms (e.g., model-based prefrontal control (Cavanagh et al., 2013) and WM (Guitart-Masip, Duzel, et al., 2014; Guitart-Masip, Economides, et al., 2014)) in controlling the Pavlovian system, they did not directly suggest active recruitment of the prefrontal WM system. First, while Cavanagh et al. speculated that midfrontal theta power could be indicative of “model-based top-down prefrontal control” (Cavanagh et al., 2013), a subsequent study by Swart et al. suggested that midfrontal theta signals could only be involved in the detection of conflict by signaling “the need for control” (Cavanagh & Frank, 2014; Swart et al., 2018), rather than being a source of direct control. Next, the IFG showed an increased BOLD response only in the “no-go” condition (Guitart-Masip et al., 2012), implying that the IFG is important for “inhibitory” motor control (i.e., as a brake (Aron et al., 2014)); it does not participate in active maintenance or representation of goal-directed behaviors including both “go” and “no-go,” which would be more closely associated with WM (Levy & Goldman-Rakic, 2000; Petrides, 2000; Rottschy et al., 2012). Finally, elevated dopamine levels should be cautiously interpreted as improvements in prefrontal WM function (Guitart-Masip, Economides, et al., 2014). While dopamine has been shown to enable successful cognitive control in the prefrontal cortex, it may have three roles: gating behaviorally relevant sensory signals; maintaining and manipulating information in WM to guide goal-directed behavior; and relaying motor information to premotor areas for action preparation (Ott & Nieder, 2019). Moreover, distinct mechanisms have been known to modulate the influence of dopamine on WM in the PFC through distinct types of dopamine receptors (Ott & Nieder, 2019). Thus, there may be several ways to interpret the observation that dopamine level (Guitart-Masip, Economides, et al., 2014) or function (Richter et al., 2021) was associated with the modulation of Pavlovian influences. Considerable research is needed to fully understand the mechanisms by which dopamine levels affect Pavlovian bias. Alternatively, the role of prefrontal WM in controlling Pavlovian bias may not require vast resources. It may only be responsible for signaling a need for control (Swart et al., 2018), promoting response inhibition (Guitart-Masip et al., 2012), or influencing subcortical areas (e.g., the striatum and subthalamic nucleus (Albrecht et al., 2016; Cools, 2016)).

Increased random choices under WM load

Another notable finding was that random choice increased under WM load. Our modeling analysis revealed that irreducible noise parameter estimates were greater in the WMGNG task than in the GNG task (**Figure 3**), suggesting increased random action selection under WM load. Further analysis using the modeling outputs revealed that participants' choices were less affected by the relative value difference between the "go" and "no-go" actions under WM load (**Figure 4A**). Moreover, analysis using the absolute difference between the two options (**Figure 4B**) revealed that the increase in accuracy became smaller as the absolute difference increased (i.e., the choice became easier). Both findings suggest that WM involvement led to an increase in random choices, regardless of value comparison and choice difficulty.

Our findings are broadly consistent with the results of previous studies concerning the role of WM and prefrontal regions in action selection and execution (Barrouillet et al., 2007; Dalley et al., 2004; Granon et al., 1994; Oberauer, 2019; Ridderinkhof et al., 2004; Ripke et al., 2012; Seo et al., 2012; Szmalec et al., 2005). In particular, several studies have demonstrated that the interruption of WM function via WM load could increase the frequency of random choices in value-based decision-making tasks (Franco-Watkins et al., 2006, 2010; Olschewski et al., 2018). Additionally, transcranial direct current stimulation, a brain stimulation method, over the left PFC led to increased random action selection during an RL task, suggesting that the prefrontal WM component influenced action selection (Turi et al., 2015). Furthermore, the importance of WM in action selection during learning tasks is supported by the indirect evidence that individual differences in WM capacity were correlated with appropriate exploratory action selection in multi-armed bandit tasks (Laureiro-Martinez et al., 2019). Overall, the reduced availability of WM resources because of WM load in our study may have compromised the participants' abilities to actively represent their current goals and actions, leading to reduced WM control over consistent choice based on value computation.

No significant neural correlates were identified with respect to the increased random choices. We assumed that random action selection would be associated with the reduced sensitivity to value difference or value comparison between the two options ("go" and "no-go") (Gläscher & O'Doherty, 2010); thus, we hypothesized that value comparison signals would decrease under WM load. Contrary to our hypothesis, no significant differences in value comparison signaling in ROIs were observed between GNG and WMGNG tasks. There are several possible explanations for this null finding. Our assumption of value sensitivity may not be the source of the random choice observed here. Alternatively, subsequent attentional lapse (Master et al., 2020; Nassar & Frank, 2016) or value-independent noise (Talmi et al., 2009) may have led to inconsistent action selection despite the presence of value comparison signals. Further research is necessary to distinguish these possibilities.

In summary, the present study has shown that WM load compromises overall learning by reducing learning speed via weakened cooperation between RL and WM; it also increases random action selection without affecting the balance between Pavlovian and instrumental systems. To our knowledge, this is the first study to investigate the neurocognitive effect of WM load during interactions between Pavlovian and instrumental systems. By investigating how learning and decision-making using different systems are altered in the presence of WM load and by linking such behaviors to their underlying neural mechanisms, this study contributes to our understanding of how distinct cognitive components interact with each other and synergistically contribute to learning. Because impairments in learning, balance among multiple systems, and action selection have been reported in various neurological and psychiatric disorders (Huys et al., 2016, 2021), our findings represent an important step toward improved understanding of various symptoms.

Materials and methods

Participants

Fifty-six adults participated in this study (34 women; 24.5±3.6 years old). All participants were healthy, right-handed; they had normal or corrected-to-normal visual acuity. They were screened prior to the experiment to exclude individuals with a history of neurological, or psychiatric illness. All participants provided written informed consent, and the study protocol was approved by the Institutional Review Board of Seoul National University.

The behavioral analysis included 49 participants (29 women; 24.3±3.3 y.o); the fMRI analysis included 44 participants (27 women; 24.2±3.3 y.o). Four participants were excluded because of technical issues; one participant was excluded because they slept during the task. Two participants were excluded because of poor performance in the 2-back task since the results in the dual-task paradigm could only be valid and interpretable when participants actually performed both tasks. The accuracy cutoff was 0.575, a value that rejects the null hypothesis that participants would randomly choose one of two options. After assessment of preprocessed image quality, five participants were excluded from the fMRI analysis because of head movements in the scanner, which can systematically alter brain signals; four out of these five were excluded because the mean framewise displacement exceeded 0.2 mm (Gu et al., 2015), while the remaining one was excluded after visual assessment of carpet plots (Power, 2017).

Experimental design and task

The experiment was performed in two blocks: one contained the original GNG task (Guitart-Masip et al., 2012) and one contained the GNG task paired with the 2-back task as a secondary task. The order of task completion was counterbalanced among participants. The

GNG and WMGNG tasks consisted of two blocks (four blocks in total); each block consisted of 60 trials. Therefore, each task contained 120 trials (240 trials in total). Participants underwent fMRI while performing the tasks for approximately 50 min, with a short (~60 s) break after each set of 60 trials. Before scanning, participants performed 20 practice trials each of GNG task and WMGNG task to help them become accustomed to the task structure and response timing. Participants received additional compensation based on their accuracy in the two tasks, along with the standard participation fee at the end of the experiment.

Orthogonalized go/no-go (GNG) task

Four trial types were implemented depending on the nature of the fractal cue (**Figure 1A**): press a button to gain a reward (go to win); press a button to avoid punishment (go to avoid losing); do not press a button to earn a reward (no-go to win); and do not press a button to avoid punishment (no-go to avoid losing). The meanings of fractal images were randomized among participants.

Each trial consisted of three phases: fractal cue presentation, response, and probabilistic outcome. **Figure 1B** illustrates the trial timeline. In each trial, participants were presented with one of four abstract fractal cues for 1000 ms. After a variable interval drawn from a uniform probability distribution within the range of 250-2000 ms, a white circle was displayed on the center of the screen for 1000 ms. When the circle appeared, participants were required to respond by pressing a button or not pressing a button. Next, the outcome was presented for 1000 ms: a green arrow pointing upwards on a ₩1000 bill indicated monetary reward, a red arrow pointing downwards on a ₩1000 bill indicated monetary punishment, and a yellow horizontal bar indicated no reward or punishment.

The outcome was probabilistic; thus, 80% correct responses and 20% incorrect responses resulted in the best outcome. Participants were instructed that the outcome would be probabilistic; for each fractal image, the correct response could be either “go” or “no-go,” and they would have to learn the correct response for each cue through trial and error. The task included 30 trials for each of the four trial types (120 trials in total). Trial types were randomly shuffled throughout the duration of the task.

Orthogonalized go/no-go + 2-back (WMGNG) task

In the WM load condition, the GNG task was accompanied by a 2-back task to induce WM load. The combined task was named the WMGNG task; each trial had 2-back response and 2-back outcome phases after the GNG task (fractal cue, response, and probabilistic outcome). Participants were required to indicated whether the cue in the current trial was identical to the cue presented in the two previous trials. For example, as shown in **Figure 1C**, the cue in the third trial differes from the cue in the first trial (two trials prior); thus, participants should respond "different" by pressing button after responding to the reinforcement learning task. In the fourth trial, they should respond "same." The positions of "SAME" and "DIFF" were randomized among participants.

Computational modeling

Baseline RL model with Pavlovian bias

We adopted a previously implemented version of an RL model (Guitart-Masip et al., 2012) that can model Pavlovian bias and choice randomness as well as learning rate. In our baseline model, we assumed no difference in parameters between the control and load conditions.

Expected values $Q(a_t, s_t)$ were calculated for each action a, "go" or "no-go", on each stimulus s (i.e., four trial types of the task) on each trial t. $Q(a_t, s_t)$ was determined by Rescorla-Wagner or delta rule updating:

$$Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \epsilon(\rho r_t - Q_{t-1}(a_t, s_t))$$

where ε is the learning rate. The learning rate (ε) is a step size of learning (Sutton & Barto, 2018) that modulates how much of the prediction error, a teaching signal, is incorporated into the value update.

Rewards, neutral outcomes, and punishments were entered in the model through $r_t \in$ {−1, 0, 1}, where ρ reflects the weighting (and effect sizes) of rewards and punishments. In all models, ρ could be different for rewards and punishments (ρ_{rew} for gain, ρ_{pun} for loss).

Action weights $W(a_t, s_t)$ were calculated from Q values, and the Pavlovian and go biases:

$$W_t(a_t, s_t) = \begin{cases} Q_t(a_t, s_t) + b + \pi V_t(s_t) & if\ a = go \\ Q_t(a_t, s_t), & else \end{cases}$$

where b was added to the value of go, while the expected value on the current state $V_t(S_t)$ was weighted by π and added to the value of go choices. $V_t(S_t)$ was computed as follows:

$$V_t(s_t) \ = \ V_{t-1}(s_t) \ + \ \epsilon(\rho r_t - V_{t-1}(s_t)).$$

If the Pavlovian bias parameter (π) is positive, it increases the action weight of "go" in the reward conditions because $V_t(S_t)$ is positive. In the punishment conditions, positive π decreases the action weight of "go" because $V_t(S_t)$ is negative.

Action probabilities were dependent on these action weights $W(a_t, s_t)$, which were passed through a squashed softmax (Sutton & Barto, 2018):

$$P(a_t, s_t) = \left[\frac{\exp[W(a_t, s_t)]}{\sum_{a'} \exp[W(a', s_t)]}\right](1 - \xi) + \frac{\xi}{2}$$

where ξ was the irreducible noise in the decision rule; it was free to vary between 0 and 1 for all models. The irreducible noise parameter explains the extent to which information about action weights is utilized in making a choice. As the irreducible noise increases, the influence of the difference between the action weights is reduced, indicating that action selection becomes random.

Additional models

To test our hypotheses regarding the effects of WM load on parameters, we constructed seven additional nested models assuming different Pavlovian biases (π), learning rate (ε), and irreducible noise (ξ) under WM load (**Table 1**). Model 1 is the baseline model. Model 2 assumed a separate Pavlovian bias parameter (π) for the WM load condition. Similarly, models 3 and 4 assumed different learning rates (ε) and irreducible noises (ξ) in the WMGNG block, respectively. To address the possibility that two of the three parameters would be affected by the WM load, we constructed three additional models with eight free parameters: model 5 with different Pavlovian bias (π) and learning rate (ε); model 6 with different Pavlovian bias (π) and irreducible noise (ξ); and model 7 with different learning rate (ε) and irreducible noise (ξ). Finally, we constructed the full model, which assumed that all of these three parameters would be affected by WM load, leading to nine free parameters.

Table 1. Free parameters of all models

Model No.	Model	# of parameters
1	$\epsilon, \rho_{rew}, \rho_{pun}, b, \pi, \xi$	6
2	$\epsilon, \rho_{rew}, \rho_{pun}, b, \pi, \xi, \pi_{wm}$	7
3	$\epsilon, \rho_{rew}, \rho_{pun}, b, \pi, \xi, \epsilon_{wm}$	7
4	$\epsilon, \rho_{rew}, \rho_{pun}, b, \pi, \xi, \xi_{wm}$	7
5	$\epsilon, \rho_{rew}, \rho_{pun}, b, \pi, \xi, \pi_{wm}, \epsilon_{wm}$	8
6	$\epsilon, \rho_{rew}, \rho_{pun}, b, \pi, \xi, \pi_{wm}, \xi_{wm}$	8
7	$\epsilon, \rho_{rew}, \rho_{pun}, b, \pi, \xi, \epsilon_{wm}, \xi_{wm}$	8
8	$\epsilon, \rho_{rew}, \rho_{pun}, b, \pi, \xi, \pi_{wm}, \epsilon_{wm}, \xi_{wm}$	9

Procedures for model fitting and model selection

Model parameters were estimated using hierarchical Bayesian analysis (HBA). Group-level distributions were assumed to be normally distributed, with mean and standard deviation parameters set as two free hyperparameters. We employed weakly informative priors to minimize the influences of those priors on the posterior distributions (Ahn et al., 2017; Kruschke, 2014). Additionally, for parameter estimation, the Matt trick was used to minimize the dependence between group-level mean and standard deviation parameters; it also facilitated the sampling process (Papaspiliopoulos et al., 2007). Moreover, bounded parameters such as learning rates and irreducible noise (∈ [0, 1]) were estimated within an unconstrained space; they were then probit-transformed to the constrained space, thus maximizing MCMC efficiency within the parameter space (Ahn et al., 2017; Wetzels et al., 2010).

We ran four independent chains with 4000 samples each, including 2000 warm-up samples (i.e., burn-in) to ensure that the parameters converged to the target distributions. Four chains were run to ensure that the posterior distributions were not dependent on initial starting points (Vehtari et al., 2019). We visually checked convergence to target distributions by observing trace plots (**Figure S1**) and computing the R statistics - a measure of convergence across chains (Gelman & Rubin, 1992). R statistics were < 1.1 for all models, indicating that the estimated parameter values converged to their target posterior distributions (**Table S1**).

Models were compared using the LOOIC, which is an information criterion calculated from the leave-one-out cross-validation (Vehtari et al., 2017). This method is used to estimate the out-of-sample predictive accuracy of a fitted Bayesian model for model comparison and selection. The LOOIC is computed using the log-likelihood evaluated from posterior distributions or simulations of the parameters. The R package loo (Vehtari et al., 2017), which provides an interface for the approximation of leave-one-out cross-validated log-likelihood, was used to estimate the LOOIC for each model. Lower LOOIC values indicated better fit.

fMRI scans: acquisition and protocol

fMRI was performed on the same scanner (Simens Tim Trio 3 Tesla) using a 32-channel head coil across all participants. A high-resolution T1-weighted anatomical scan of the whole brain resolution was also acquired for each participant (TR = 2300ms, TE = 2.36ms, FOV = 256mm,1mm×1mm×1mm) to enable spatial localization and normalization. The participant's head was positioned with foam pads to limit head movement during acquisition. Functional data was acquired using echo-planar imaging (EPI) in four scanning sessions containing 64 slices (TR = 1500ms, TE = 30ms, FOV = 256mm, 2.3mm × 2.3mm × 2.3mm). For the GNG task, functional imaging data were acquired in two separate 277-volume runs, each lasting about 7.5 min. For the WMGNG task, data were acquired in two separate 357-volume runs, each lasting about 9.5 min.

fMRI scans: general linear models

Preprocessing was performed using fMRIPrep 20.2.0 (Esteban et al., 2018, 2019; RRID:SCR_016216), which is based on Nipype 1.5.1 (K. Gorgolewski et al., 2011; K. J. Gorgolewski et al., 2018; RRID:SCR_002502). Details of preprocessing with fMRIPrep are provided in Supplementary Material. Subsequently, images were smoothed using a 3D Gaussian kernel (8mm FWHM) to adjust for anatomical differences among participants. BOLD-signal image analysis was then performed using SPM12 [http://www.fil.ion.ucl.ac.uk/spm/] running on MATLAB v9.5.0.1067069(R2018b).

We built participant-specific GLMs, including all runs – two runs for the GNG block and two runs for the WMGNG block – and calculated contrasts to compare the two blocks at the individual level. The first-level model included six movement regressors to control the movement-related artifacts as nuisance regressors. Linear contrasts at each voxel were used to obtain participant-specific estimates for each effect. These estimates were entered into group-level analyses, with participants regarded as random effects, using a one-sample t-test against a contrast value of 0 at each voxel. The group-level model included covariates for gender, age, and the task order. For all GLM analyses, we conducted ROI analysis; the results were corrected for multiple comparisons using small volume correction (SVC) within ROIs.

GLM1 (Hypothesis 1): GLM1 was used to test hypothesis 1: RPE signaling in the striatum would be increased under WM load. Therefore, GLM was implemented by the model-based fMRI approach and included the following regressors: (1) cue onset of “go to win” trials, (2) cue onset of “no-go to win” trials, (3) cue onset of “go to avoid losing” trials, (4) cue onset of “no-go to avoid losing” trials, (5) target onset of “go” trials, (6) target onset of “no-go” trials, (7) outcome onset, (8) outcome onset parametrically modulated by the trial-by-trial RPEs, and (9) wait onset (i.e., inter-trial interval). The regressor of interest was “RPE”; we compared the main effect of RPE between two tasks ($\text{RPE(8)}_{\text{WMGNG}}$ - $\text{RPE(8)}_{\text{GNG}}$). RPE regressors were calculated by subtracting the expected values (Q) from the outcome for each trial. Here, the outcome was the product of feedback multiplied by reward/punishment sensitivity. ROI was the striatum, which is widely known to process RPE (Chase et al., 2015; Garrison et al., 2013; J. P. O'Doherty et al., 2003).

GLM2 (Hypothesis 2): GLM2 was used to test hypothesis 2: neural responses associated with Pavlovian bias would be increased under WM load. Specifically, the GLM examined whether the difference between the anticipatory response to fractal cues in Pavlovian-congruent trials and Pavlovian-incongruent trials was greater in the WMGNG task than in the GNG task in regions associated with Pavlovian bias. Therefore, GLM included the following regressors: (1) cue onset of "go to win" trials, (2) cue onset of "no-go to win" trials, (3) cue onset of "go to avoid losing" trials, (4) cue onset of "no-go to avoid losing" trials, (5) target onset of "go" trials, (6) target onset of "no-go" trials, (7) outcome onset of win trials, (8) outcome onset of neutral trials, (9) outcome onset of loss trials, (10) wait onset (i.e., inter-trial interval). We compared the main effect of Pavlovian bias (Pavlovian-congruent trials - Pavlovian-incongruent trials) between two tasks ($[(1) + (4) - ((2) + (3))]_{WMGNG}$ - $[(1) + (4) ((2) + (3))]_{GNG}$). ROIs included the striatum and SN/VTA. The striatum ROI was constructed by combining the AAL3 definitions of bilateral caudate, putamen, olfactory bulb, and nucleus accumbens. Furthermore, the SN/VTA was constructed by combining the AAL3 definitions of bilateral SN and VTA.

GLM3 (Hypothesis 3): GLM3 was used to test hypothesis 3: value comparison signals would decrease under WM load. GLM3 was also implemented with a model-based fMRI approach: (1) cue onset of all trials, (2) cue onset parametrically modulated by the trial-by-trial decision values (W_{chosen} - $W_{unchosen}$), (3) target onset of "go" trials, (4) target onset of "no-go" trials, (5) outcome onset, and (6) wait onset (i.e., inter-trial interval). Decision value regressors were calculated by subtracting the action weights of the unchosen option ($W_{unchosen}$) from the action weights of the chosen option (W_{chosen}). We then compared the main effect of decision value between two blocks ($(2)_{WMGNG}$-$(2)_{GNG}$). ROIs for GLM3 included the vmPFC, which was suggested as a region that encodes the relative chosen value (W_{chosen} - $W_{unchosen}$) (Boorman et al., 2009; S. W. Lee et al., 2014). Here, ROI masks were created by drawing a sphere with a diameter of 10 mm around the peak voxel reported in the previous studies ([-6,48,-8] for vmPFC (Boorman et al., 2009)).

PPI analysis: In addition to GLMs, we used PPI analysis to test whether WM load led to reduced cooperation between WM and RL systems for learning (Collins, Ciullo, et al., 2017; Collins & Frank, 2018) by using PPI analysis. Here, to examine differences between the two blocks in terms of functional coupling between the prefrontal areas and the area computing RPE after choices, we performed PPI analysis using the gPPI toolbox (McLaren et al., 2012); the physiological variable was the time course of the striatum, and the psychological variable was the effect of WM load during the anticipation phase. As a seed region (i.e., a physiological variable), the cluster striatum ROI (peak MNI space coordinates x = 13, y = 14, z = -3) was derived from the results of GLM2, which revealed stronger RPE signaling in the WMGNG task than in the GNG task. The entire time series throughout the experiment was extracted from each participant in the striatum ROI. To create the PPI regressor, these normalized time series were multiplied by task condition vectors for the anticipation phase, which consisted of the cue representation and fixation phases as in GLM1. A GLM with PPI regressors of the seed region was thus generated together with movement regressors. The effects of PPI for each participant were estimated in the individual-level GLM; the parameter estimates represented the extent to which activity in each voxel was correlated with activity in the striatum during the anticipation phase. The contrast was constructed by subtracting activity during the anticipation phase in the GNG task from activity in the WMGNG task (WMGNG vs. GNG in the anticipation phase). Individual contrast images for functional connectivity were then computed and entered into one-sample t-tests in a group-level GLM together with nuisance covariates (i.e., gender, age, and task order). Whole-brain cluster correction was applied for PPI analysis.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Science, ICT, & Future Planning (NRF-2018R1C1B300731315 and NRF-2018R1A4A102589113); the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021M3E5D2A0102249311); and the Creative-Pioneering Reserachers Program through Seoul National University to W.-Y.A.

References

Adams, R. A., Moutoussis, M., Nour, M. M., Dahoun, T., Lewis, D., Illingworth, B., Veronese, M., Mathys, C., Boer, L. de, Guitart-Masip, M., Friston, K. J., Howes, O. D., & Roiser, J. P. (2020). Variability in Action Selection Relates to Striatal Dopamine 2/3 Receptor Availability in Humans: A PET Neuroimaging Study Using Reinforcement Learning and Active Inference Models. *Cerebral Cortex*, *30*(6), 3573–3589. https://doi.org/10.1093/cercor/bhz327

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry*, *1*, 24–57. https://doi.org/10.1162/cpsy_a_00002

Albrecht, M. A., Waltz, J. A., Cavanagh, J. F., Frank, M. J., & Gold, J. M. (2016). Reduction of Pavlovian Bias in Schizophrenia: Enhanced Effects in Clozapine-Administered Patients. *PLoS ONE*, *11*(4), e0152781. https://doi.org/10.1371/journal.pone.0152781

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (20*14*). Inhibition and the right inferior frontal cortex: one decade on. *Trends in Cognitive Sciences*, *18*(4), 177–185.

Baddeley, A. D., & Hitch, G. (1974). Working Memory. *Psychology of Learning and Motivation*, *8*, 47–89. https://doi.org/10.1016/s0079-7421(08)60452-1

Ballard, I. C., Murty, V. P., Carter, R. M., MacInnes, J. J., Huettel, S. A., & Adcock, R. A. (2011). Dorsolateral Prefrontal Cortex Drives Mesolimbic Dopaminergic Regions to Initiate Motivated Behavior. *Journal of Neuroscience*, *31*(*28*), 10340–10346. https://doi.org/10.1523/jneurosci.0895-11.2011

Balleine, B. W., & O'doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69.

Barbey, A. K., Koenigs, M., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human working memory. Cortex, *49*(5), 1195–1205.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and Cognitive Load in Working Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 570–585. https://doi.org/10.1037/0278-7393.33.3.570

Betts, M. J., Richter, A., Boer, L. de, Tegelbeckers, J., Perosa, V., Baumann, V., Chowdhury, R., Dolan, R. J., Seidenbecher, C., Schott, B. H., Düzel, E., Guitart-Masip, M., Krauel, K. (2020). Learning in anticipation of reward and punishment: perspectives across the human lifespan. *Neurobiology of Aging*, *96*, 49–57. https://doi.org/10.1016/j.neurobiolaging.2020.08.011

Boer, L. de, Axelsson, J., Chowdhury, R., Riklund, K., Dolan, R. J., Nyberg, L., Bäckman, L., & Guitart-Masip, M. (2018). Dorsal striatal dopamine D1 receptor availability predicts an instrumental bias in action learning. *Proceedings of the National Academy of Sciences*, *116*(1), 201816704. https://doi.org/10.1073/pnas.1816704116

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, *62*(5), 733–743. https://doi.org/10.1016/j.neuron.2009.05.014

Bornstein, A. M., & Daw, N. D. (2011). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology*, *21*(3), 374–380.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *16*(11), 681–684. https://doi.org/10.1037/h0040090

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., & Frank, M. J. (2013). Frontal Theta Overrides Pavlovian Learning Biases. *The Journal of Neuroscience*, *33*(19), 8541–8548. https://doi.org/10.1523/jneurosci.5754-12.2013

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. Trends in Cognitive Sciences, 18(8), 414–421. https://doi.org/10.1016/j.tics.2014.04.012

Chase, H. W., Kumar, P., Eickhoff, S. B., & Dombrovski, A. Y. (2015). Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 435–459. https://doi.org/10.3758/s13415-015-0338-7

Chowdhury, R., Guitart-Masip, M., Lambert, C., Dolan, R. J., & Düzel, E. (2013). Structural integrity of the substantia nigra and subthalamic nucleus predicts flexibility of instrumental learning in older-age individuals. Neurobiology of Aging, 34(10), 2261–2270.

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. https://doi.org/10.1162/jocn_a_01238

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. *Biological Psychiatry*, *82*(6), 431–439. https://doi.org/10.1016/j.biopsych.2017.05.017

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working memory load strengthens reward prediction errors. Journal of Neuroscience, 37(16), 4332–4342.

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proceedings of the National Academy of Sciences, *115*(10), 201720963. https://doi.org/10.1073/pnas.1720963115

Collins, A. G. E., & Shenhav, A. (2021). Advances in modeling learning and decision-making in neuroscience. Neuropsychopharmacology, 1–15. https://doi.org/10.1038/s41386-021-01126-y

Cools, R. (2016). The costs and benefits of brain dopamine for cognitive control. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(5), 317–329.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. Trends in Cognitive Sciences, 7(9), 415–423.

Dalley, J. W., Cardinal, R. N., & Robbins, T. W. (2004). Prefrontal executive and cognitive functions in rodents: neural and neurochemical substrates. *Neuroscience & Biobehavioral Reviews*, *28*(7), 771–784.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

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

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, *22*(3), 213–219.

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. Neural Networks, 19(8), 1153–1160. https://doi.org/10.1016/j.neunet.2006.03.002

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67–78.

Dickinson, A., & Balleine, B. (2002). *The role of learning in the operation of motivational systems.*

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. Neuron, 80(2), 312–325.

Dorfman, H. M., & Gershman, S. J. (2019). Controllability governs the balance between Pavlovian and instrumental action selection. *Nature Communications*, *10*(1), 5826. https://doi.org/10.1038/s41467-019-13737-7

Economides, M., Guitart-Masip, M., Kurth-Nelson, Z., & Dolan, R. J. (2014). Anterior Cingulate Cortex Instigates Adaptive Switches in Choice by Integrating Immediate and Delayed Components of Value in Ventromedial Prefrontal Cortex. The Journal of Neuroscience, 34(9), 3340–3349. https://doi.org/10.1523/jneurosci.4313-13.2014

Ereira, S., Pujol, M., Guitart-Masip, M., Dolan, R. J., & Kurth-Nelson, Z. (2021). Overcoming Pavlovian bias in semantic space. *Scientific Reports*, *11*(1), 3416. https://doi.org/10.1038/s41598-021-82889-8

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Kent, J. D. andGoncalves, DuPre, E., Sitek, K. R., Gomez, D. E. P., Lurie, D. J., Ye, Z., Poldrack, R. A., & Gorgolewski, K. J. (2018). fMRIPrep. *Software*. https://doi.org/10.5281/zenodo.852659

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., & others. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116.

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

Franco-Watkins, A. M., Pashler, H., & Rickard, T. C. (2006). Does Working Memory Load Lead to Greater Impulsivity? Commentary on Hinson, Jameson, and Whitney (2003). Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(2), 443–447. https://doi.org/10.1037/0278-7393.32.2.443

Franco-Watkins, A. M., Rickard, T. C., & Pashler, H. (2010). Taxing Executive Processes Does Not Necessarily Increase Impulsive Decision Making. Experimental Psychology, 57(3), 193–201. https://doi.org/10.1027/1618-3169/a000024

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science, *306*(5703), 1940–1943.

Friston, K., Buechel, C., Fink, G., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage*, *6*(3), 218–229.

Funahashi, S. (2006). Prefrontal cortex and working memory processes. Neuroscience, *139*(1), 251–261.

Funahashi, S., & Kubota, K. (1994). Working memory and prefrontal cortex. *Neuroscience Research*, *21*(1), 1–11.

Garbusow, M., Schad, D. J., Sebold, M., Friedel, E., Bernhardt, N., Koch, S. P., Steinacher, B., Kathmann, N., Geurts, D. E., Sommer, C., & others. (2016). Pavlovian-to-instrumental transfer effects in the nucleus accumbens relate to relapse in alcohol dependence. *Addiction Biology*, *21*(3), 719–731.

Garbusow, M., Schad, D. J., Sommer, C., Jünger, E., Sebold, M., Friedel, E., Wendt, J., Kathmann, N., Schlagenhauf, F., Zimmermann, U. S., Heinz, A., Huys, Q. J. M., & Rapp, M. A. (2014). Pavlovian-to-Instrumental Transfer in Alcohol Dependence: A Pilot Study. *Neuropsychobiology*, *70*(2), 111–121. https://doi.org/10.1159/000*363*507

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. Neuroscience & Biobehavioral Reviews, 37(7), 1297–1310. https://doi.org/10.1016/j.neubiorev.2013.03.023

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. Wiley Interdisciplinary Reviews: Cognitive Science, 1(4), 501–510.

Glasner, S. V., Overmier, J. B., & Balleine, B. W. (2005). The role of Pavlovian cues in alcohol seeking in dependent and nondependent rats. *Journal of Studies on Alcohol*, *66*(1), 53–61.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C., Yvernault, B., Salo, T., Jordan, K., Goncalves, M., Waskom, M., Clark, D., Wong, J., Loney, F., … Ghosh, S. (2018). Nipype. Software. https://doi.org/10.5281/zenodo.596855

Granon, S., Vidal, C., Thinus-Blanc, C., Changeux, J.-P., & Poucet, B. (1994). Working memory, response selection, and effortful processing in rats with medial prefrontal lesions. Behavioral Neuroscience, *108*(5), 8*83*.

Gu, S., Satterthwaite, T. D., Medaglia, J. D., Yang, M., Gur, R. E., Gur, R. C., & Bassett, D. S. (2015). Emergence of system roles in normative neurodevelopment. Proceedings of the National Academy of Sciences, 112(44), 13681–13686.

Guitart-Masip, M., Duzel, E., Dolan, R., & Dayan, P. (2014). Action versus valence in decision making. Trends in Cognitive Sciences, 18(4), 194–202. https://doi.org/10.1016/j.tics.2014.01.003

Guitart-Masip, M., Economides, M., Huys, Q. J. M., Frank, M. J., Chowdhury, R., Duzel, E., Dayan, P., & Dolan, R. J. (2014). Differential, but not opponent, effects of l-DOPA and citalopram on action learning with reward and punishment. *Psychopharmacology*, *231*(5), 955–966. https://doi.org/10.1007/s00213-013-3313-4

Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J.

(2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage*, *62*(1), *154*–166. https://doi.org/10.1016/j.neuroimage.2012.04.024

Hershberger, W. A. (1986). An approach through the looking-glass. *Animal Learning & Behavior*, *14*(4), 443–451.

Hiser, J., & Koenigs, M. (2018). The Multifaceted Role of the Ventromedial Prefrontal Cortex in Emotion, Decision Making, Social Cognition, and Psychopathology. Biological Psychiatry, 83(8), 638–647. https://doi.org/10.1016/j.biopsych.2017.10.030

Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. Neuropsychopharmacology, 46(1), 3–19. https://doi.org/10.1038/s41386-020-0746-4

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience, 19(3), 404–413. https://doi.org/10.1038/nn.4238

Kahnt, T., Heinzle, J., Park, S. Q., & Haynes, J.-D. (2011). Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. Neuroimage, *56*(2), 709–715.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5), e1002055.

Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., & Guillot, A. (2005). Actor--Critic models of reinforcement learning in the basal ganglia: from natural to artificial rats. *Adaptive Behavior*, *13*(2), 131–148.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.*

Laureiro-Martinez, D., Brusoni, S., Tata, A., & Zollo, M. (2019). The Manager's Notepad: Working Memory, Exploration, and Performance. *Journal of Management Studies*, *56*(8), 1655–1682. https://doi.org/10.1111/joms.12528

Lee, D., & Seo, H. (2007). Mechanisms of reinforcement learning and decision making in the primate dorsolateral prefrontal cortex. *Annals of the New York Academy of Sciences*, *1104*(1), 108–122.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. Neuron, 81(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028

Levy, R., & Goldman-Rakic, P. S. (2000). Segregation of working memory functions within the dorsolateral prefrontal cortex. *Executive Control and the Frontal Lobe: Current Issues*, 23–32.

Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2011). The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention. The Journal of Neuroscience, 31(37), 13214–13223. https://doi.org/10.1523/jneurosci.1246-11.2011

Lüscher, C., Robbins, T. W., & Everitt, B. J. (2020). The transition to compulsion in addiction. *Nature Reviews Neuroscience*, *21*(5), 247–263. https://doi.org/10.1038/s41583-020-0289-z

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Clarendon Press Oxford.

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. E. (2020). Disentangling the systems contributing to changes in learning during adolescence. *Developmental Cognitive Neuroscience*, *41*, 100732. https://doi.org/10.1016/j.dcn.2019.100732

McDougle, S. D., & Collins, A. G. E. (2020). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, 1–20. https://doi.org/10.3758/s13423-020-01774-z

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. NeuroImage, *61*(4), 1277–1286. https://doi.org/10.1016/j.neuroimage.2012.03.068

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience, 16(5), 1936–1947.

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, *11*, 49–54.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–1*54*. https://doi.org/10.1016/j.jmp.2008.12.005

Oberauer, K. (2019). Working Memory and Attention – A Conceptual Analysis and Review. *Journal of Cognition*, *2*(1), 36. https://doi.org/10.5334/joc.58

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science, *304*(5669), 452–454.

O'Doherty, J. P. (2011). Contributions of the ventromedial prefrontal cortex to goal-directed action selection. Annals of the New York Academy of Sciences, *1239*(1), 118–129. https://doi.org/10.1111/j.1749-6632.2011.06290.x

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. Neuron, *38*(2), 329–337. https://doi.org/10.1016/s0896-6273(03)00169-7

Olschewski, S., Rieskamp, J., & Scheibehenne, B. (2018). Taxing Cognitive Capacities Reduces Choice Consistency Rather Than Preference: A Model-Based Test. *Journal of Experimental Psychology: General*, *147*(4), 462–484. https://doi.org/10.1037/xge0000403

Ott, T., & Nieder, A. (2019). Dopamine and Cognitive Control in Prefrontal Cortex. Trends in Cognitive Sciences, 23(3), 213–234. https://doi.org/10.1016/j.tics.2018.12.006

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*(5), 751–761.

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. Statistical Science, 59–73.

Park, S. Q., Kahnt, T., Beck, A., Cohen, M. X., Dolan, R. J., Wrase, J., & Heinz, A. (2010). Prefrontal cortex fails to learn from reward prediction errors in alcohol dependence. Journal of Neuroscience, 30(22), 7749–7753.

Pasupathy, A., & Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. Nature, *433*(7028), 873–876.

Perosa, V., Boer, L. de, Ziegler, G., Apostolova, I., Buchert, R., Metzger, C., Amthauer, H., Guitart-Masip, M., Düzel, E., & Betts, M. J. (2020). The Role of the Striatum in Learning to Orthogonalize Action and Valence: A Combined PET and 7 T MRI Aging Study. Cerebral Cortex, 30(5), 3340–3351. https://doi.org/10.1093/cercor/bhz313

Petrides, M. (2000). The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research*, *133*(1), 44–54.

Power, J. D. (2017). A simple but useful way to assess fMRI scan qualities. NeuroImage, 154(NeuroImage 10 1999), 150–158. https://doi.org/10.1016/j.neuroimage.2016.08.009

Pujara, M. S., Philippi, C. L., Motzkin, J. C., Baskaya, M. K., & Koenigs, M. (2016). Ventromedial Prefrontal Cortex Damage Is Associated with Decreased Ventral Striatum Volume and Response to Reward. The Journal of Neuroscience, 36(18), 5047–5054. https://doi.org/10.1523/jneurosci.4236-15.2016

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience, 9(7), 545–556. https://doi.org/10.1038/nrn2357

Richter, A., Boer, L. de, Guitart-Masip, M., Behnisch, G., Seidenbecher, C. I., & Schott, B. H. (2021). Motivational learning biases are differentially modulated by genetic determinants of striatal and prefrontal dopamine function. *Journal of Neural Transmission*, 1–16. https://doi.org/10.1007/s00702-021-02382-4

Richter, A., Guitart-Masip, M., Barman, A., Libeau, C., Behnisch, G., Czerney, S., Schanze, D., Assmann, A., Klein, M., Düzel, E., Zenker, M., Seidenbecher, C. I., & Schott, B. H. (2014). Valenced action/inhibition learning in humans is modulated by a genetic variant linked to dopamine D2 receptor expression. *Frontiers in Systems Neuroscience*, *8*, 140. https://doi.org/10.3389/fnsys.2014.00140

Ridderinkhof, K. R., Wildenberg, W. P. V. D., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition*, *56*(2), 129–140.

Ripke, S., Hübner, T., Mennigen, E., Müller, K. U., Rodehacke, S., Schmidt, D., Jacob, M. J., & Smolka, M. N. (2012). Reward processing and intertemporal decision making in adults and adolescents: The role of impulsivity and decision consistency. Brain Research, *1478*, 36–47. https://doi.org/10.1016/j.brainres.2012.08.034

Rmus, M., McDougle, S. D., & Collins, A. G. (2021). The role of executive function in shaping reinforcement learning. Current Opinion in Behavioral Sciences, 38, 66–73. https://doi.org/10.1016/j.cobeha.2020.10.003

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., Fox, P. T., & Eickhoff, S. B. (2012). Modelling neural correlates of working memory: A coordinate-based meta-analysis. NeuroImage, *60*(1), 830–846. https://doi.org/10.1016/j.neuroimage.2011.11.050

Schultz, W. (1997). Dopamine neurons and their role in reward mechanisms. Current Opinion in Neurobiology, 7(2), 191–197.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. Science, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Seo, M., Lee, E., & Averbeck, B. B. (2012). Action selection and action value in frontal-striatal circuits. Neuron, 74(5), 947–960.

Smith, D. V., Hayden, B. Y., Truong, T.-K., Song, A. W., Platt, M. L., & Huettel, S. A. (2010). Distinct Value Signals in Anterior and Posterior Ventromedial Prefrontal Cortex. Journal of Neuroscience, 30(7), 2490–2495. https://doi.org/10.1523/jneurosci.3319-09.2010

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Swart, J. C., Frank, M. J., Määttä, J. I., Jensen, O., Cools, R., & Ouden, H. E. M. den. (2018). Frontal network dynamics reflect neurocomputational mechanisms for reducing maladaptive biases in motivated action. *PLoS Biology*, *16*(10), e2005979. https://doi.org/10.1371/journal.pbio.2005979

Szmalec, A., Vandierendonck, A., & Kemps, E. (2005). Response selection involves executive control: Evidence from the selective interference paradigm. *Memory & Cognition*, *33*(3), 531–541.

Talmi, D., Dayan, P., Kiebel, S. J., Frith, C. D., & Dolan, R. J. (2009). How humans integrate the prospects of pain and reward during choice. Journal of Neuroscience, 29(46), 14617–14626.

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008). Calculating consequences: brain systems that encode the causal effects of actions. Journal of Neuroscience, 28(26), 6750–6755.

Tsujimoto, S., & Sawaguchi, T. (2004). Neuronal representation of response--outcome in the primate prefrontal cortex. Cerebral Cortex, 14(1), 47–55.

Turi, Z., Mittner, M., Opitz, A., Popkes, M., Paulus, W., & Antal, A. (2015). Transcranial direct current stimulation over the left prefrontal cortex increases randomness of choice in instrumental learning. Cortex, 63, 145–154.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. Journal of Neuroscience, 27(15), 4019–4026.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved $\widehatR$ for assessing convergence of MCMC. *ArXiv Preprint ArXiv:1903.08008*.

Wallis, J. D., & Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. European Journal of Neuroscience, 18(7), 2069–2081. https://doi.org/10.1046/j.1460-9568.2003.02922.x

Wasserman, E. A., Franklin, S. R., & Hearst, E. (1974). Pavlovian appetitive contingencies and approach versus withdrawal to conditioned stimuli in pigeons. *Journal of Comparative and Physiological Psychology*, *86*(4), 616–627. https://doi.org/10.1037/h0036171

Wasserman, E. A., & Miller, R. R. (1997). What's elementary about associative learning? *Annual Review of Psychology*, *48*(1), 573–607.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. Journal of Mathematical Psychology, 54(1), 14–27.

Williams, D. R., & Williams, H. (1969). AUTO-MAINTENANCE IN THE PIGEON: SUSTAINED PECKING DESPITE CONTINGENT NON-REINFORCEMENT 2. *Journal of the Experimental Analysis of Behavior*, *12*(4), 511–520.

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. Journal of Cognitive Neuroscience, 34(4), 551–568. https://doi.org/10.1162/jocn_a_01808