

Research Articles: Behavioral/Cognitive

Neural index of reinforcement learning predicts improved stimulus-response retention under high working memory load

<https://doi.org/10.1523/JNEUROSCI.1274-22.2023>

Cite as: J. Neurosci 2023; 10.1523/JNEUROSCI.1274-22.2023

Received: 29 June 2022

Revised: 19 January 2023

Accepted: 20 February 2023

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

28 **Abstract**

29 Human learning and decision making is supported by multiple systems operating in parallel. Recent
30 studies isolating the contributions of reinforcement learning (RL) and working memory (WM) have
31 revealed a trade-off between the two. An interactive WM-RL computational model predicts that while
32 high WM load slows behavioral acquisition, it also induces larger prediction errors in the RL system
33 that enhance robustness and retention of learned behaviors. Here we tested this account by
34 parametrically manipulating WM load during RL in conjunction with EEG, in both male and female
35 participants, and administered two surprise memory tests. We further leveraged single trial decoding
36 of EEG signatures of RL and WM to determine whether their interaction predicted robust retention.
37 Consistent with the model, behavioral learning was slower for associations acquired under higher load
38 but showed parametrically improved future retention. This paradoxical result was mirrored by EEG
39 indices of RL, which were strengthened under higher WM loads and predictive of more robust future
40 behavioral retention of learned stimulus-response contingencies. We further tested whether stress
41 alters the ability to shift between the two systems strategically to maximize immediate learning versus
42 retention of information and found that induced stress had only a limited effect on this trade-off. The
43 present results offer a deeper understanding of the cooperative interaction between WM and RL and
44 show that relying on WM can benefit the rapid acquisition of choice behavior during learning but
45 impairs retention.

46

47

48

49

50

51

52

53

54

55

56 **Significance statement**

57 Successful learning is achieved by the joint contribution of the dopaminergic reinforcement learning
58 (RL) system and working memory (WM). The cooperative WMRL model was productive in
59 improving our understanding of the interplay between the two systems during learning, demonstrating
60 that reliance on RL computations is modulated by WM load. However, the role of WM/RL systems in
61 the retention of learned stimulus-response associations remained unestablished. Our results show that
62 increased neural signatures of learning, indicative of greater RL computation, under high WM load
63 also predicted better stimulus-response retention. This result supports a trade-off between the two
64 systems, where degraded WM increases RL processing which improves retention. Notably, we show
65 that this cooperative interplay remains largely unaffected by acute stress.

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84 **Introduction**

85 Everyday behavior, like selecting what to wear and what to eat, involves reinforcement
86 learning (RL). Canonical RL models incrementally accumulate expected values of stimulus-action
87 pairings over the course of multiple experiences. While this RL system learns rather slowly and
88 incrementally, it can be augmented by the joint support of working memory (WM), especially when
89 learning new arbitrary contingencies (Yoo & Collins, 2021). WM enables fast learning by robustly
90 maintaining, in an accessible form, the representations of relevant stimulus-action associations to
91 support ongoing processing such as value-based learning and decision-making. However, when WM
92 capacity is exceeded, it suffers from interference, causing relevant representations to be lost or
93 corrupted (Oberauer et al., 2016). Indeed, while the WM system is beneficial for supporting early
94 learning, its contribution to successful learning is constrained by limited capacity (Collins & Frank,
95 2012). On the other hand, the incremental RL system has a much broader capacity and is more robust
96 as long as the reward contingencies remain stable. Previous studies have thus shown a transition from
97 capacity- and delay-sensitive WM to RL over the course of learning (Collins & Frank, 2012; 2018).

98 Moreover, recent studies examining the joint contributions of WM and RL to learning have
99 suggested that these systems are not modular, but rather interactive (Collins, 2018; Collins & Frank,
100 2018; Collins et al., 2017a,b). fMRI and EEG studies provided support for a cooperative interaction:
101 when stimulus-reward information is stored in WM, neural indices of reward prediction errors (RPEs)
102 are reduced (Collins et al., 2017a; Collins & Frank, 2018). Conversely, RPEs were larger under high
103 load, leading to accelerated “neural learning curves” putatively indicative of more robust RL (despite
104 slowed behavioral learning due to degraded WM). This dissociation suggested that while a high WM
105 load slows learning, it might also improve retention, due to accumulative RPEs that reinforce the RL
106 system. Supporting this prediction, in the surprise test phase, participants showed better retention
107 performance for stimulus-response contingencies and their reward values when they had been learned
108 under higher compared to lower WM demands (Collins et al., 2017b; Collins, 2018; Wimmer &
109 Poldrack, 2020). However, two major limitations remained from this prior work.

110 First, the previous study showing enhanced retention of stimulus-response associations had
111 only tested low and high WM conditions (Collins, 2018), with only subtle albeit significant differences

112 in performance (around 5% difference between set size 3 vs. 6). We thus parametrically manipulated
113 WM demands (Collins et al., 2017b) to test the prediction that retention performance of stimulus-
114 response associations would scale monotonically as a function of increased WM demand, despite
115 monotonically slowed learning in these conditions. Second, while the neural and behavioral findings
116 have been documented on their own, it has not yet been established whether cooperative neural
117 interactions within WM/RL systems during learning are predictive of future retention. Moreover, it is
118 unclear whether neural RL learning curves reflect reward expectations, or whether they reflect learned
119 policies (as predicted by Q learning vs. actor-critic; Jaskir & Frank 2022; Li & Daw 2011). We thus
120 sought to test these relationships directly by recording EEG during learning and then administering
121 two retention tests. The EEG measures of RL were used to assess whether the neural RL measure is
122 predictive of participants' ability to retrieve learned reward expectations and/or the retention of
123 stimulus-response contingencies.

124 As a secondary aim, we also examined the impacts of acute stress on RL and WM processes. There is
125 accumulating evidence, across various domains of learning, that acute stress reduces goal-directed
126 decision making and alters prefrontal cortex functioning (see review by Arnsten, 2009), thereby
127 promoting a shift from cognitively demanding but flexible systems towards simpler but more rigid
128 systems (e.g., Wirz, et al., 2018; Kim et al., 2001; Schwabe & Wolf, 2009; Vogel, Fernández, Joëls, &
129 Schwabe, 2016; Meier, Staesina, & Schwabe, 2022). We thus tested whether stress could reduce
130 WM's ability to effectively guide learning and instead enhance the relative contribution of RL
131 processing.

132 **Methods**

133 *Participants*

134 Eighty-six healthy volunteers (43 women, age 18-34, mean = 24.56, SD = 3.84) participated in
135 this experiment. All participants were right-handed, had normal or corrected-to-normal vision, and
136 were screened for possible EEG contraindications. Individuals with a current medical condition,
137 medication intake, or lifetime history of any neurological or psychiatric disorders were excluded from

138 participation. All participants provided written informed consent before the beginning of testing and
139 received moderate monetary compensation. The study protocol was approved by the ethics committee
140 of the Faculty of Psychology and Human Movement Sciences at the University of Hamburg.

141

142 *Experimental procedure*

143 *Learning task*

144 Interactions of RL and WM were tested using the RLWM task (Collins 2018, Collins & Frank,
145 2012; 2018), programmed in MATLAB using the Psychophysics Toolbox. In this task (see Fig. 1A),
146 each trial started with a presentation of a stimulus in the center of the screen, on a black background
147 and participants had to learn which of the three actions (key presses A1, A2, A3) to select based on
148 trial-by-trial reward feedback. Stimulus presentation and response time were limited to 1.4 sec.
149 Incorrect choices led to feedback 0, while correct choices led to reward, (reward was 1 or 2 points
150 fixed with the probability of 0.2, 0.5, or 0.8). Stimulus probability assignment was counterbalanced
151 within participants to ensure equal overall value of different set sizes (see below) and motor actions.
152 The key press was followed by audio-visual feedback (the word “Win!” with an ascending tone or the
153 word “Loss!” with a descending tone). If participants did not respond within 1.4 sec, the message
154 “Too slow!” appeared. Feedback was presented for 0.4 – 0.8 sec and followed by a fixation cross for
155 0.4 – 0.8 before the next trial started.

156 To manipulate WM demands, the number of stimulus-action contingencies to be learned
157 varied by block between 1 to 5 (denoted as ns), with new stimuli set presented at each new block (e.g.,
158 colors, fruits, or animals). There were four blocks in which set size=2, two blocks in which set size=4,
159 and three block in which set size=1, 3, 5 for a total of 15 blocks and 645 trials. Within a block, each
160 stimulus was presented 15 times. 108 stimuli were pseudo-randomized and 43 stimuli were presented
161 for each participant. Stimulus category assignment to block set size was counterbalanced across
162 subjects. Block order was also counterbalanced with the exception of set size=1 which served as
163 control (block numbers 8 and 14 were saved for set size=1).

164 The following instructions were given to participants: “In this experiment, you will see an
165 image on the screen. You need to respond to each image by pressing one of the three buttons on the
166 Gamepad: 1, 2, or 3 with your right hand. Your goal is to figure out which button makes you win for
167 each image. You will have a few seconds to respond. Please respond to every image as quickly and
168 accurately as possible. If you do not respond, the trial will be counted as a loss. If you select the
169 correct button, you will gain points. You can gain either 1 or 2 points designated as "\$" or "\$\$". Some
170 images will give you more points for correct answers on average than other images. You can only gain
171 points when you select the correct button for each image. At the beginning of each block, you will be
172 shown the set of images for that block. Take some time to identify them correctly. Note the following
173 important rules: There is ONLY ONE correct response for each image. One response button MAY be
174 correct for multiple images, or not be correct for any image. Within each block, the correct response
175 for each image will not change”.

176

177 *Test phase*

178 After the learning phase, participants completed two surprise test phases (Fig 1 B, C). The first
179 was a reward retention test that has been used in earlier studies (e.g., Collins et al., 2017b). *The reward*
180 *retention test* was designed to test whether expected values are learned by default since several
181 previous studies showed that participants can select actions based on their relative expected values at
182 the transfer phase even when they only had to learn which item was best (e.g., Frank et al, 2007;
183 Palminteri et al, 2015). In this phase, on each trial participants were requested to select the more
184 rewarding stimulus from a pair of stimuli that had each been encountered during the learning phase.
185 All stimuli that were used in the learning phase were presented in the test phase at least once. The two
186 stimuli were pseudo-randomly selected to sample across all possible combinations of set sizes, blocks
187 and probabilities. To ensure no new learning at this phase, participants did not receive any feedback on
188 their responses. Note that in this test, participants could not leverage information they had learned
189 about which response to select (the ‘policy’); instead they had to use novel response mappings to
190 simply indicate which stimulus had been more rewarded. Participants’ ability to select the more

218 of stimulus-response associations to be learned (denoted as *setSize*), and the number of intervening
219 trials since the last time the stimulus was presented and a correct response was made (denoted as
220 *delay*) reflecting WM interference or maintenance time in WM. The RL variable is the total number of
221 previous correct responses for a stimulus (denoted as *Pcor*). Participants and all the predictors were
222 selected as random variables.

223 *Behavioral analysis of the reward retention test*

224 To quantify the possible effect of expected value learning under different WM loads, we
225 analyzed test performance (the proportion of selecting the right vs left stimulus) with general mixed
226 effect regression on trial-by-trial data from 86 participants, as a function of six variables: value
227 difference (denoted as *delta_Q*; is positive when the right stimulus had higher value and negative
228 when the left stimulus had higher value), mean Q value of the stimulus pair (denoted as *mean value*
229 (*Q*)), mean set size of the stimulus pair (denoted as *mean_setSize*), the difference in set size (denoted
230 as *delta_setSize*; is positive when the right stimulus was learned in higher set size), *block* (the block
231 number in which they were learned, indicating how recently it was learned), and *perseveration* (binary
232 coding of repetitions in response, repeat/switch). Participants, the effect of value difference (*delta_Q*),
233 and the effect of set size difference (*delta_setSize*) were entered as random variables.

234 *Behavioral analysis of the reward retention test together with EEG RL index*

235 We ran a new regression model on the reward retention test data (including only the 77
236 participants that had EEG data), adding the difference in the EEG RL index between the pair of stimuli
237 at choice. Because the neural RL index (see a detailed description of this measure below) could have
238 both positive and negative values all the predictors that were calculated as difference scores were
239 taken as absolute scores and the model predicted performance accuracy (proportion of choosing the
240 higher value stimulus). Test performance accuracy was analyzed as a function of: The absolute model
241 estimated value difference between the right and left stimulus (*abs_delta_Q*); the absolute difference
242 in the EEG RL index between the right and left stimulus (*abs_delta_EEG_RL*); the mean value
243 (estimated from the model) of the stimulus pair (*mean Q value*); the mean set size of the stimulus pair

244 (*mean set size*); the absolute difference in the block number where the right and left stimulus were
245 learned (*abs_delta_block*); response bias towards the previously selected response (*perseveration*;
246 binary coding of repetitions in response). Participants, the effect of value difference (*abs_delta_Q*),
247 and the effect of EEG RL index difference (*abs_delta_EEG_RL*) were entered as random variables.

248 *Behavioral analysis of the stimulus-response retention test*

249 In a general mixed-effect regression analysis we tested accuracy for correctly recalling the
250 response associated with a presented stimulus learned during the training phase as a function of *set*
251 *size* (the set size block in which they were learned), *block* (the block number in which they were
252 learned, indicating how recently it was learned) and *model Q* (the model estimated Q value of each
253 stimulus calculated as the average Q value of the final 6 iterations during learning) and *perseveration*
254 (the tendency to repeat the response selected in the previous trial at test coded as 1 for repeat and 0 for
255 switch). The interactions between set size and model Q value, set size and block, and between set size
256 and perseveration were also added as predictors. Participants and the interaction between model Q and
257 set size were entered as random variables.

258 *Behavioral analysis of the stimulus-response retention test together with EEG RL index*

259 We ran the same regression model on the stimulus-response retention test data as before
260 (including only the 77 participants that had EEG data), adding two new predictors: the average EEG
261 RL index for each stimulus-response association (see a detailed description of this measure below) and
262 the interaction between EEG RL index and set size. Participants, the interaction between model Q and
263 set size, and the interaction between EEG RL index and set size were entered as random variables.

264

265 *Electroencephalogram (EEG) recording and processing*

266 During the learning task, participants were seated approximately 80 cm from the monitor in an
267 electrically shielded and sound attenuated cabin. EEG was recorded using a 64-channel BioSemi
268 ActiveTwo system (BioSemi, Amsterdam, The Netherlands) with sintered Ag-AgCl electrodes

269 organized according to the 10-20 system. The sampling rate was 2048 Hz. The signal was digitized
270 using a 24-bit A/D converter. Additional electrodes were placed at the left and right mastoids,
271 approximately 1 cm above and below the orbital ridge of each eye and at the outer canthi of the eyes
272 for measurement of eye movements. The EEG data were re-referenced offline to a common average.
273 Electrode impedances were kept below 30 k Ω . EEG and EOG were amplified with a low cut-off
274 frequency of 0.53 Hz (=0.3 s time constant).

275 The EEG data were processed using EEGLAB (Delorme and Makeig, 2004) and ERPLAB
276 (Lopez-Calderon and Luck, 2014). The continuous EEG was bandpass-filtered offline between 0.5–20
277 Hz and down-sampled to 125 Hz, then it was segmented into epochs ranging from 500 ms pre-
278 stimulus up to 3000 ms post-stimulus. The epoched data were visually inspected and those containing
279 large artifacts due to facial electromyographic (EMG) activity or other artifacts, except for eye blinks
280 were manually removed (e.g., large fluctuations in voltage across several electrodes that were in an
281 order magnitude above neighbouring activity). Independent components analysis (ICA) was next
282 conducted only on the 64 scalp electrodes using EEGLAB's runica algorithm. Components containing
283 blink or oculomotor artifacts, were subtracted from the data resulting in an average of 1.6 components
284 removed per participant (ranging between 0 to 3 components). Finally, the epoched data was subjected
285 to automatic bad-electrodes and artifact detection algorithm (100 μ V voltage threshold with a moving
286 window width of 200ms and a 100ms window step) which was followed by manual verification. Bad-
287 electrodes were interpolated and trials containing large artifacts were removed. Nine participants were
288 removed from all the reported EEG analyses due to a high EEG artifact rate (>40% in one or more of
289 the conditions) resulting in 77 participants that were used in the EEG analysis.

290

291 *Data processing for behavior and EEG regression analysis*

292 Omission trials, trials with very fast RT (<200ms), and trials before the first correct response
293 was made were excluded from all analyses. Setting the delay and Pcor variables to have 1 as their
294 lowest level was done to insure an interpretable analysis of these variables (Collins & Frank, 2012).
295 The delay predictor (the number of trials since the stimulus was presented and a correct response was
296 made) used in the regression analyses was inverse transformed (-1/delay) to avoid the disproportion

297 effect of very large but rare delays (when a correct response was given early in the block but was then
298 followed by several error responses for that stimulus).

299 *Modeling*

300 RL and WM contributions to participants' choices were estimated with the previously
301 developed RLWM computational model (the model described below is identical to that used in Collins
302 and Frank, 2018; see more details described in the original paper). The RLWM is a mixture of a
303 standard RL module with a delta rule and a WM module that has perfect memory for information that
304 is within its limited capacity and is sensitive to delay (reflecting memory decay and interference from
305 other intervening stimuli). For each stimulus-action association, the RL module estimates the expected
306 value ("Q") and updates those values incrementally on every trial as a function of the reinforcement
307 history. This computation is complemented by the WM module where information in the capacity-
308 limited WM feeds into RL expectations, thereby affecting RL prediction errors and learning (see Fig.
309 2).

310 **Basic RL module:** To maintain consistency with prior studies with this task and model, and to
311 keep the model as simple as possible, we use Q learning for the model-free algorithm, but an actor
312 critic could also have been used (there are multiple options to capture incremental model-free RL,
313 including methods that learn expected values for each choice and select on that basis (a canonical
314 instance is Q learning and is often used in human studies) as well as methods that learn to directly
315 optimize the policy (a canonical variant is an actor-critic model). Both classes of models similarly
316 predict behavioral adjustment in RL tasks and specific designs are needed to distinguish between them
317 (e.g., Gold et al, 2012; Geana et al 2021). The main goal here is to simply summarize the incremental
318 RL process as distinct from the WM process.

319 Reward values were coded as 0 or 1 for correct or incorrect (model fits are not improved if
320 using 1 vs 2 points in the Q learning system, and behavioral learning curves are similar for stimuli that
321 yield higher or lower probability of 2 points; Collins et al., 2017b). For each stimulus s and action a

322 association, the RL module estimates the expected reward value Q and updates those values
 323 incrementally on every trial:

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha \times \delta_t$$

324 The Q value was updated as a function of the learning rate α (reflecting how fast reward
 325 expectations are updated) and the reward prediction error δ , calculated as the difference between the
 326 observe reward, R_t and the expected reward, Q_t at each trial: $\delta_t = R_t - Q_t$.

327 Choices were probabilistically determined using a softmax choice policy:

$$p(a|s) = \exp(\beta Q(s, a)) / \sum_i (\exp(\beta Q(s, a_i)))$$

328 Here, β is the inverse temperature determining the degree to which differences in Q values are
 329 translated into more deterministic choices, and the sum is over the three possible actions. Q-values
 330 were initialized to $1/n_A$, where $n_A = 3$ is the number of actions (i.e., the prior that any action is correct
 331 is 1/3).

332 **WM module:** This module updates stimulus-action-outcome associations in a single trial. It
 333 assumes that stimulus-action-outcome information, when encoded and maintained in WM, could serve
 334 to update reward expectation rapidly and accurately (i.e., perfect retention of the previous trial's
 335 information). When not limited by capacity and decay (see below), the WM module is therefore
 336 represented by a Q learning system with a learning rate of 1 ($\alpha = 1$).

337 *Decay:* To account for potential forgetting on each trial due to delay or WM interference, we included
 338 a decay parameter ϕ ($0 < \phi < 1$) which pulls the estimates of Q values toward their initial value, [$Q_0 =$
 339 $1/n_A$, number of actions $n_A = 3$].

$$Q \leftarrow Q + \phi(Q_0 - Q)$$

340 Only the WM module was subject to forgetting (decay parameter ϕ_{WM}), to capture WM's well
 341 documented short-term stability, in contrast to RL's robustness.

342 *WM contributes to choice:* Because WM is capacity limited, only K stimulus and action associations
 343 can be remembered. A constraint factor reflects the *a priori* probability that the item was stored in
 344 WM: $w_{WM}(0) = P_0(WM) = K/n_s$ (i.e., the set size in the current block relative to capacity K) and
 345 implies that the maximal use of WM policy relative to RL policy depends on the probability that an
 346 item is stored in WM. This probability is then scaled by ρ ($0 < \rho < 1$), the participant's overall reliance
 347 of WM vs RL (where higher values reflect greater confidence in WM).

$$w_{WM}(0) = \rho * \min(1, K/n_s)$$

348 **Cooperative model:** While the original model (Collins & Frank, 2012) assumed independent
 349 RL and WM modules that compete to guide behavior, our more recent work suggests that WM
 350 expectations influence RL updating (Collins & Frank, 2018). Thus, WM contributes part of the reward
 351 expectation for the RL model, according to the equation: $\delta_t = R_t - [w_{WM} \times Q_{WM} + (1 - w_{WM}) \times$
 352 $Q_{RL}]$, where w_{WM} is the weighting parameter (the degree to which WM is weighted relative to RL,
 353 which is stronger in low set sizes), and Q_{WM} is the expected reward from the WM module. This RPE
 354 is then used to update the RL Q value: $Q_{t+1} = Q_t + \alpha \times \delta_t$

355 This interactive computation of RL forms the basis of the simulated predictions shown in Figure 2.
 356 Nevertheless, as explained in Collins and Frank (2018), we test these predictions by fitting models in
 357 which RL and WM modules are independent (independence is assumed in the original models, which
 358 still provide good fits to the data, because when information is within WM, WM dominates updating
 359 and contributes to rapid learning curves, and hence the interactive models' smaller RPEs and RL Q
 360 values for small set sizes are not influential on behavioral accuracy during learning; however, this
 361 model makes differential predictions for neural learning curves and future retention). We then assess
 362 systematic deviations from independence informed by these simulations (e.g, neural Q learning curves
 363 should grow more rapidly in high than low set sizes; Fig. 2).

364 ----- figure 2 is here -----

365

366 *Data processing for univariate EEG analysis*

367 To extract the neural correlates in the EEG signal of conditions of interest we employed a
368 mass univariate approach (Collins & Frank, 2018). A multiple regression analysis was conducted for
369 each participant, in which the EEG amplitude at each electrode site and time point was predicted by
370 the conditions of interest: set-size (number of stimulus-response-outcome associations given in a
371 block), model-derived RL expected value (denoted as Q), delay (number of trials since this stimulus
372 was presented and a correct response was given) and the interaction of these three regressors, while
373 controlling for other factors like reaction time (log-transformed) and trial number within block.
374 Furthermore, the EEG signal was reduced to a selected window of -100 to +700 ms around stimulus
375 onset, and was baseline corrected from -100 to 0 ms before the onset of the stimulus. To account for
376 remaining noise in the EEG data, the EEG signal (at each time point and electrode) was z-scored
377 across all trials and so were all the predictors before they were entered to the robust multilinear
378 regression analysis (Collins & Frank, 2018).

379 *Corrected ERPs*

380 To plot corrected ERPs, we computed the predicted voltage using the multiple-regression
381 model described above while setting a single regressor to 0 (set size, delay, expected Q value, or
382 reaction time); we subtracted this predicted voltage from the true voltage (for every electrode and time
383 point within each trial), leaving only the fixed effect, the variance explained by that regressor, and the
384 residual noise of the regression model. ERPs were computed as the average corrected voltage from all
385 trials that belong to the same level of condition. Note that the array of expected Q values was divided
386 to 4 quartiles and trials within each quartile were averaged for plotting ERPs.

387 *Trial-by-trial similarity index of WM and RL*

388 As explained above, a multiple regression analysis was conducted for each participant, in
389 which the EEG amplitude at each electrode site and time point was predicted by the conditions of
390 interest (set size, delay, RL expected value, and their interactions). We used the previously identified
391 analysis method (Collins & Frank, 2018; Rac-Lubashevsky & Frank 2021) to identify spatiotemporal

392 clusters (masks) of the three main predictors in the GLM (set-size, delay, and model-derived RL
393 expected value). Specifically, we tested the significance of each time point at each electrode across
394 participants against 0 using only trials with correct responses.
395 We then used cluster-mass correction by permutation testing with custom written Matlab scripts.
396 Cluster-based test statistics were calculated by taking the sum of the t-values within a spatiotemporal
397 cluster of points that exceeded the $P = 0.001$ threshold for a t-test significance level. This was repeated
398 1000 times, generating a distribution of maximum cluster-mass statistics under the null hypothesis.
399 Only clusters with greater t-value sum than the maximum cluster-mass obtained with 95% chance
400 permutations were considered significant. We then assessed each trial's neural similarity to the
401 spatiotemporal mask by computing the dot product between the activity in the individual trial (voltage
402 maps of electrode \times time) and the identified masks (t-value maps of electrode \times time). This
403 computation produced a trial-level similarity measure intended to assess the trial-wise experienced
404 WM load and delay effects, as well as trial-wise RL contributions.

405 The EEG RL index predictor used in the general mixed-effect regression analyses of both test
406 phases was calculated by averaging the EEG RL index in the final 6 iterations of each stimulus. This
407 was done for each stimulus-response association within each participant.

408 *Stress manipulation*

409 All testing took place in the morning between 8am and noon. Upon their arrival in the lab,
410 participants' baseline measures of blood pressure and salivary cortisol were taken. Afterwards,
411 participants were prepared for the EEG and completed the mood questionnaire MDBF (Steyer, et al.,
412 1994) that measures subjective mood on the scales negative vs. elevated mood, calmness vs.
413 restlessness, and wakefulness vs. tiredness, before and after the treatment as well as after the learning
414 task. 42 participants underwent the Socially-evaluated Cold Pressor Test (SECPT; Schwabe et al.,
415 2008) and 44 participants were assigned the warm water control condition. The SECPT is a
416 standardized stress protocol in experimental stress research that combines physiological and
417 psychosocial stress elements and has been shown to result in robust stress responses (Schwabe &
418 Schächinger, 2018). During the SECPT, participants in the stress group immersed their right hand for

419 three minutes in ice water (0-2°C), while being videotaped and evaluated by a non-reinforcing, cold
420 experimenter. In the control condition, participants immersed their hands in warm water (35-37°C),
421 without being videotaped or evaluated by an experimenter. About 25 minutes after the treatment,
422 participants received the learning task instructions and completed a brief training session after which
423 they completed the learning task and the test phases 1 and 2. In total, the experiment lasted about 130
424 minutes.

425 Results

426 In line with previous findings in this task (e.g., Collins et al. 2017b), our data demonstrated
427 separable contributions of RL and WM systems to performance. The contribution of incremental RL
428 was observed as the proportion of correct responses increased with the progress in the block (Fig. 3A)
429 and with the increase in reward history (p_{cor} : $\beta=.67$, $SE=.05$, $z(46926)=13.17$, $p<.001$). WM
430 contributions were observed as learning was strongly affected by set size with a greater proportion of
431 correct responses in low set sizes than in high set sizes ($set\ size$: $\beta=-.28$, $SE=.05$, $z(46926)=-5.39$,
432 $p<.001$). Learning curves were more gradual in higher set sizes than in low set sizes (Fig. 3A; and
433 slower Fig. 3B). Moreover, performance decreased with increasing delay in larger set sizes ($delay \times$
434 ns , $\beta=-.09$, $SE=.05$, $z(46926)=-2.59$, $p=.009$; Fig. 3C). These relative contributions of WM decreased
435 with learning as the detrimental effect of delay attenuated with the increase of accumulated rewards
436 ($ns \times P_{cor}$: $\beta=.13$, $SE=.04$, $z(46926)=3.35$, $p<.001$; $delay \times P_{cor}$: $\beta=.34$, $SE=.04$, $z(46926)=9.17$,
437 $p<.001$; $ns \times Delay \times P_{cor}$, $\beta=.20$, $SE=.03$, $z(46926)=6.37$, $p<.001$; Fig. 3D-3E), reflecting a
438 transition from WM to RL. Together these results confirm the cooperative interaction of early WM
439 contributions that diminish as RL becomes more dominant.

440 --- Figure 3 is here---

441 Behavioral Performance: Reward Retention Test

442 Results replicated previous findings in this phase (Collins et al, 2017b). Participants were
443 more likely to select the stimulus for which they had been rewarded more often during learning as a
444 function of the difference between the number of rewards experienced for these stimuli (δ_{Q} :

445 $\beta=.41$, $SE=.04$, $z(19796)=9.76$, $p<.001$). Moreover, also replicating previous findings, this value
446 discrimination effect was enhanced when stimulus values were learned under higher set sizes rather
447 than under lower set sizes ($mean_setSize \times delta_Q$: $\beta=.11$, $SE=.02$, $z(19796)=6.04$, $p<.001$). For
448 display purposes, the median split in the absolute $delta_Q$ score is shown as high and low-value
449 differences (see Fig. 4A). Furthermore, participants were generally less likely to select the stimulus
450 learned under a higher set size than under a low set size ($delta_setSize$, $\beta=-.69$, $SE=.09$, $z(19796)=-$
451 7.61 , $p<.001$), an effect previously attributed to participants learning a cost of mental effort in a high
452 set size (Collins et al 2017b). There was no effect for the difference in the block in which the item
453 values were learned nor was the set size effect modulated by block number ($p >.82$). We also
454 controlled for response perseveration; no significant tendency was observed for repeating the same
455 response used in the previous trial ($p >.69$).

456

457

--- Figure 4 is here---

458

459 *Behavioral Performance: Stimulus-response retention test*

460 Supporting the key model prediction that retention of stimulus-response associations should
461 improve as load increases, we observed better recall performance for associations learned under high
462 rather than low set sizes ($set\ size$: $\beta=.84$, $SE=.05$, $z(11894)=15.83$, $p<.001$). And, indeed this effect
463 was parametric, with substantially better performance as set size increased (see Fig. 4B-C). This effect
464 is particularly striking given that performance is parametrically worse for the higher set size items
465 during learning (compare Fig. 3A and Fig. 4C). Not surprisingly, recall accuracy in the test phase was
466 positively predicted by the estimated Q value of the probed stimulus-response association ($model\ Q$:
467 $\beta=.27$, $SE=.04$, $z(11894)=6.97$, $p<.001$), that is, associations that were learned better were also better
468 remembered. Importantly, this effect grew when the set size was high ($model\ Q \times set\ size$: $\beta=.15$,
469 $SE=.04$, $z(11894)=3.64$, $p<.001$; see Fig. 4B). Recall accuracy was also subject to the influence of
470 recency as associations learned during more recent than early blocks were also recalled more
471 accurately ($block$: $\beta=.22$, $SE=.03$, $z(11894)=8.61$, $p<.001$). This recency effect increased for

499

--- Figure 6 is here ---

500

501 We next assessed the core prediction that the neural RL index is related to future retention, and
502 more specifically, the cooperative model prediction that the speeded neural RL curves in high set sizes
503 are related to better retention of learned contingencies. Notably, while this prediction did not hold for
504 the reward retention phase ($abs_delta_EEG_RL: p=.65$; $mean_setSize \times abs_delta_EEG_RL: p=.61$;
505 Fig 4D), it was clearly borne out for the stimulus-response retention phase ($EEG_RL: \beta=.23$,
506 $z(10613)=4.51, p<.001$; Fig 4E). Stimuli that had been associated with a larger EEG RL index during
507 learning were associated with better recall of the associated response at test; this effect held even when
508 controlling for the non-neural predictors (which replicated the prior analysis). Figure 4E shows that a
509 high EEG RL index (by median split) was predictive of better retention performance at test. The
510 finding that the neural index of RL is related to policy retention but not reward retention is relevant for
511 models that dissociate whether model-free RL in the brain encodes expected values or policies (see
512 model method section and Discussion). Note that a slightly different regression model was used for
513 testing the neural RL index effect on the reward retention test performance than the behaviour model
514 used previously (see Method section for more detail). Nevertheless, the key behavior results were
515 replicated in this analysis as performance increased with the increase in the absolute value differences
516 ($abs_delta_Q: \beta=.31, SE=.03, z(17743)=8.82, p<.001$) and while this effect was not further modulated
517 by set size ($mean_setSize \times abs_delta_Q, p=.63$), performance accuracy did improve with set size
518 ($mean_setSize: \beta=.07, SE=.02, z(17743)=3.23, p=.001$; see Fig 4D).

519

520 *Acute stress modulation of RL and WM interaction*521 *Manipulation check*

522 Subjective, autonomic and endocrine data indicated that the stress induction by the SECPT
523 was successful. The SECPT was rated as significantly more unpleasant, stressful, and painful than the
524 warm water control procedure: [more difficult, $t(84) = 9.941, p <.001, d = 2.14$; more unpleasant, $t(84)$
525 $= 9.088, p <.001, d = 1.96$; more stressful, $t(84) = 7.72, p <.001, d = 1.66$; and more painful $t(84) =$
526 $11.42, p <.001, d = 2.46$; see rating reports in Table 1]. Furthermore, we observed significant

527 Treatment-by-Time interactions for subjective stress ratings [negative mood: $F_{2,164} = 10.53, p < .001,$
 528 $\eta_g^2 = .02$; restlessness: $F_{2,164} = 9.47, p < .001, \eta_g^2 = .02$] and autonomic arousal measures [systolic
 529 blood pressure (SBP): $F_{4,336} = 26.22, p < .001, \eta_g^2 = .06$; diastolic blood pressure (DBP): $F_{4,336} =$
 530 $26.99, p < .001, \eta_g^2 = .09$; and heart rate: $F_{3,252} = 10.70, p < .001, \eta_g^2 = .02$]. As expected, these
 531 autonomic responses returned relatively quickly to baseline after the treatment (see Fig.6). The stress
 532 and no-stress control groups did not differ in any of the autonomic arousal measures pre-treatment (all
 533 p-values>.07).

534 *--- Figure 6 is here ---*

535 *--- Table 1 is here ---*

536 Salivary cortisol (sCORT) responses were assessed by running ANOVA with Time (T1, T2,
 537 T3, T4) as the within-subject factor and Treatment (SECPT vs. warm water control group) as the
 538 between-subject factor. We observed a significant effect for Time ($F_{3,234} = 28.53, p < .001, \eta_p^2 = .27$)
 539 but not for Treatment ($F_{1,78} = 3.03, p = .08, \eta_p^2 = .04$). An expected Treatment \times Time interaction was
 540 observed ($F_{3,234} = 6.97, p < .001, \eta_p^2 = .08$), with the stress group displaying greater sCORT levels
 541 immediately before the learning task (23 min post-treatment) [$t(78) = 2.80, p = .006, d = 0.63$] but
 542 only marginal difference was observed at half time during learning task (50 min post-treatment) [$t(78)$
 543 $= 1.90, p = .06, d = 0.43$]. No difference in sCORT levels was observed at baseline [$t(78) = 0.61, p =$
 544 $.54$] nor at the end of the learning task (80 min post-treatment) [$t(78) = 0.11, p = .91$], suggesting that
 545 stress-induced cortisol elevations gradually decreased during the learning task (Fig. 10). Note that 6
 546 participants were excluded from the cortisol analysis because they did not provide sufficient saliva for
 547 analysis.
 548

549

550 *Learning Phase performance by stress group*

551 To test the hypothesis that acute stress may reduce WM's ability to effectively guide learning
 552 thereby weakening the relative contribution of WM in the training phase in the stress group compared
 553 to the control group, we ran the same general mixed-effect regression model on trial-by-trial training
 554 data from 86 participants but added stress group as a factor (42 participants in the stress group and 44

555 participants in the control group). This analysis revealed that learning by set size interaction was
556 modulated by stress ($pcor \times set\ size \times stress_group: \beta=-.20, SE=.08, z(46926)=-2.60, p=.009$) and so
557 was the learning by delay interaction ($pcor \times delay \times stress_group: \beta=.22, SE=.07, z(46926)=3.04,$
558 $p=.002$). To understand the nature of these interactions we ran two follow-up analyses using the same
559 general mixed-effect regression model on trial-by-trial training data, separately in the control (N=44)
560 and the stress group (N=42). These analyses showed that learning curves were additive to the set size
561 effect in the stress group ($pcor \times set\ size: p=.74$) but not in the control group ($pcor \times set\ size: \beta=.22,$
562 $SE=.05, z(24031)=4.30, p<.001$) which showed a greater drop in performance during high set sizes
563 (see Fig. 7A-B). The attenuated delay effect with learning was significant for both the stress group
564 ($pcor \times delay: \beta=.47, SE=.05, z(22895)=8.41, p<.001$) and the control group ($pcor \times delay: \beta=.23,$
565 $SE=.05, z(24031)=4.74, p<.001$; see Fig. 7C-D).

566 *--- Figure 7 is here ---*

567

568 *Reward Retention Test performance by stress group*

569 To test the hypothesis that acute stress may reduce WM's ability to effectively guide learning
570 thereby strengthening RL conurbations during the training phase and leading to better retention of
571 learned information in the stress group compared to the control group, we ran the same general mixed-
572 effect regression model on trial-by-trial reward retention test data from 86 participants but added stress
573 group as a factor (42 participants in the stress group and 44 participants in the control group) and
574 analyzed test performance (the proportion of selecting the right vs left stimulus). This analysis
575 replicated the results of the behavior analysis without the group factor. No effect of stress was
576 observed ($p>.15$; Fig 7E).

577 *Stimulus-response retention test performance by stress group*

578 To test the hypothesis that acute stress may reduce WM's ability to effectively guide learning
579 thereby strengthening RL conurbations during the training phase and leading to better retention of
580 learned information in the stress group compared to the control group, we ran the same general mixed-

581 effect regression model on trial-by-trial stimulus-response retention test data from 86 participants but
582 added stress group as a factor (42 participants in the stress group and 44 participants in the control
583 group) and analyzed test performance. This analysis revealed that the effect of set size on recall
584 accuracy of stimulus-response associations interacted with stress (*set size* × *stress_group*: $\beta=.22$,
585 $SE=.10$, $z(11894)=2.30$, $p=.02$; Fig. 7F) but follow up analysis on each group separately showed
586 significant effect of set size on recall accuracy in both the control group ($\beta=.72$, $SE=.07$,
587 $z(6129)=10.72$, $p<.001$) and the stress group ($\beta=.95$, $SE=.08$, $z(5765)=11.76$, $p<.001$).

588 Discussion

589 Taken together, our findings provide insight into the intricate interplay between WM and RL
590 during learning, and its opposing influences on acquisition vs. retention of stimulus-response
591 associations. A recent study proposed a cooperative WMRL model, whereby RPEs in the RL system
592 are not only computed relative to RL expected values but are also modulated by expectations held in
593 WM (Collins & Frank, 2018). This model accounted for fMRI and EEG findings in which neural
594 RPEs were diminished for smaller WM loads (Collins et al., 2017; Collins & Frank, 2018). Moreover,
595 this model accounted for findings that on a given trial, larger neural indices of WM expectations were
596 predictive of subsequent RPEs during the outcome, even within a given set size (Collins & Frank,
597 2018). This model led to a key prediction that enhanced RL processes under high WM load would
598 support more robust retention of learned association, despite the substantially slower acquisition.
599 Preliminary behavioral evidence for such a behavioral prediction had been reported by Collins (2018),
600 who showed enhanced retention of items learned in set size 6 compared to set size 3. However, that
601 study did not employ neural recordings and thus did not test whether the neural WMRL interaction
602 was the underlying mechanism for these effects. Here we provide several lines of evidence in support
603 of this claim.

604 First, our behavioral and EEG results replicated key findings in the RLWM task and in the
605 subsequent memory tests. In the learning task, we observed worse acquisition with increasing set size
606 and with delays between successive stimulus presentations, but as learning progressed (with the
607 increase in reward history) the negative effect of delay in high set sizes diminished considerably. This

608 observation further supports the model prediction that RL dominates over WM with the accumulation
609 of rewards over time. Second, at the neural level, we also replicated findings in which neural RL
610 indices preceded the cognitively costly WM process during stimulus processing (Collins and Frank,
611 2018). Moreover, we found robust evidence that EEG signals of RL increased more rapidly across
612 trials under high than low load (Fig. 4F), a key prediction of the cooperative model (Fig. 2), even
613 though behavioral learning was slower in these conditions.

614 Importantly, we observed that associations learned under higher WM load had increasingly
615 higher recall accuracy in the stimulus-response retention test (Fig. 4C). This result extends the
616 previously reported retention benefit of associations learned under high compared to low set sizes
617 (Collins, 2018). We showed that this effect is parametric across five levels of WM load, and moreover
618 that the greatest retention deficits occurred for the very lowest set sizes in which participants could
619 easily learn the task purely via WM. Furthermore, we replicated previous results in the reward
620 retention test (Collins et al., 2017) and demonstrated that participants have differential sensitivity to
621 the proportion of trials in which they were rewarded for either of the stimuli and this effect grew with
622 set size.

623 Finally, to gain a better understanding of the mechanism responsible for the benefits in both
624 retention tests, we leveraged a within-trial neural indexing approach of EEG dynamics. We showed
625 that neural indices of RL during acquisition were predictive of subsequent retention in the stimulus-
626 response retention, even after controlling for set size. This result supports the key model prediction
627 that RL processes during learning, which are stronger under high WM load, are responsible for
628 increasing policy retention, when WM is no longer available. In contrast, neural indices of RL were
629 not predictive of performance in the reward retention test.

630 This result supports theoretical and empirical studies suggesting that model-free learning in
631 the brain (especially the corticostriatal system) directly learns a stimulus-response policy using
632 prediction errors from another system (“actor-critic”; Collins & Frank 2014; Jaskir & Frank 2022;
633 Klein et al 2017). By this account, the “actor” selecting policies would have no direct access to
634 experienced reward values, but only the propensity for a specific response for each of them.
635 Participants could plausibly access their “critic” values for each stimulus and compare them in the

636 reward retention phase, but they would not have had to do so during learning. Indeed, participants
637 show above chance performance in such discriminations, but only subtly (accuracy rises up to 60% at
638 best); in contrast, accuracy in the stimulus-response retention test, which directly assesses what the
639 actor would have learned, is far superior (roughly 80% for the higher set sizes), despite being tested
640 with further delays since learning.

641 For most simple RL tasks, these two classes of model-free RL algorithms (those that focus on
642 learning expected values and the actor-critic), are largely indistinguishable as they both predict that an
643 agent progressively chooses those actions that maximize reward. However, several theoretical and
644 empirical studies suggest that the basic RL system in humans satisfies predictions of an actor-critic in
645 behavior, imaging, and in theoretical models of corticostriatal contributions to RL (Collins & Frank
646 2014; Jaskir & Frank 2022; Li & Daw 2011; Klein et al 2017; Gold et al 2012; Geana et al., 2021).
647 Moreover, the model fits here did not improve if we allowed the Q learning agent to learn the
648 difference between 2 vs 1 point, and instead suggested that participants learned to simply maximize
649 task performance, which effectively makes Q learning equivalent to an actor-critic at the level of task
650 performance. Nevertheless, a Q learner would, at minimum, learn the reward value of a stimulus in
651 terms of the percentage of times they were correct (i.e., whether they got 1 or 2 points vs 0). Yet, the
652 EEG marker of RL is still not related to performance in reward retention test even when correct
653 performance there would be counted as simply choosing the stimulus that had yielded higher
654 proportion of correct responses. While our neural RL index cannot distinguish between an EEG metric
655 of “Q values”, or “actor weights”, the findings that it only predicts performance in the stimulus-
656 response test provides initial evidence supporting the actor interpretation where the neural RL index
657 reflects the policy rather than its reward value.

658 While we focussed mainly on how the RLWM mechanism informs retention, we also tested
659 whether the interaction between RL and WM can be modulated by acute stress. Stress is known to
660 have a major impact on learning and decision-making processes (Cremer et al., 2021; Raio, et al.,
661 2017; Starcke & Brand, 2012). Previous work had shown that acute stress alters prefrontal cortex
662 functioning thus impairing executive control over cognition (e.g., cognitive inhibition, task switching,
663 working memory maintenance; Bogdanov & Schwabe, 2016; Brown et al., 2020; Hamilton &

664 Brigman, 2015; Goldfarb et al., 2017; Plessow et al., 2012; Schwabe & Wolf, 2011; Schwabe, et al.,
665 2011; Vogel et al., 2016). On the other hand, acute stress was also shown to increase striatal dopamine
666 activity (Vaessen et al., 2015) leading to better working-memory updating (Goldfarb et al., 2017) and
667 improving executive control over motor actions (i.e., response inhibition; Leong and Packard, 2014;
668 Schwabe & Wolf, 2012). We, therefore, predicted that stress would affect the WM vs. RL trade-off
669 such that it will impede WM's contribution to learning and will instead enhance the relative
670 contribution of RL computations. Current results did not confirm this hypothesis as only subtle
671 differences were observed between the stress and control groups during the learning task and at the
672 tests.

673 It is possible that the 25 minutes' delay between the stressor and the beginning of the learning
674 task hindered the stress response on behavior as it was previously suggested that both noradrenaline
675 and cortisol levels need to be elevated in order for stress to affect WM performance (Roosendaal, et
676 al., 2006; Barsegyan et al., 2010; Elzinga & Roelofs, 2005). Another intriguing possibility is that
677 individuals with higher WM capacity were more resilient against cognitive impairments induced by
678 stress and were also less biased toward habitual decision-making (Cremer et al., 2021; Otto et al.,
679 2013; Quaedflieg et al., 2019). Future work should test directly the specific effect of stress on WM and
680 RL interactions while taking into account participants' WM capacity as a factor.

681 To conclude, our results contribute to a better understanding of the coupled mechanism of
682 WM and RL that can dynamically shift between relying more on the effortful but fast and reliable WM
683 system or the slow, more error-prone RL system that has retention benefits. We reported trial-by-trial
684 evidence in the neural signal for this trade-off during learning and showed that greater reliance on the
685 RL system when WM is degraded (i.e., when WM load is high) predicted better memory retention of
686 learned stimulus-response associations. An intriguing possibility that remains to be tested is that the
687 shift between the two systems is strategic and can be modulated by one's preference or ability to
688 maximize immediate learning vs retention. However, it remains to be seen if clinical populations with
689 impairments in one or both systems of WM and RL, might alter the flexible shifting between the two
690 systems, possibly biasing the use of one system more than the other even when it is less advantageous.

691 **References**

- 692 1. Arnsten, A. F. (2009). Stress signalling pathways that impair prefrontal cortex structure and
693 function. *Nature reviews neuroscience*, *10*, 410-422.
- 694 2. Barsegyan, A., Mackenzie, S. M., Kurose, B. D., McGaugh, J. L., & Roozendaal, B. (2010).
695 Glucocorticoids in the prefrontal cortex enhance memory consolidation and impair working
696 memory by a common neural mechanism. *Proceedings of the National Academy of
697 Sciences*, *107*, 16655-16660.
- 698 3. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models
699 using lme4. arXiv preprint arXiv:1406.5823. doi: <https://doi.org/10.48550/arXiv.1406.5823>
700
- 701 4. Bogdanov, M., & Schwabe, L. (2016). Transcranial stimulation of the dorsolateral prefrontal
702 cortex prevents stress-induced working memory deficits. *Journal of Neuroscience*, *36*, 1429-
703 1437.
- 704 5. Brown, T. I., Gagnon, S. A., & Wagner, A. D. (2020). Stress disrupts human hippocampal-
705 prefrontal function during prospective spatial navigation and hinders flexible behavior.
706 *Current Biology*, *30*, 1-13.
- 707 6. Carvalheiro, J., Conceição, V. A., Mesquita, A., & Seara-Cardoso, A. (2021). Acute stress
708 impairs reward learning in men. *Brain and Cognition*, *147*, 105657.
- 709 7. Collins, C. J., Yi, F., Dayuha, R., Duong, P., Horslen, S., Camarata, M., ... and Hahn, S. H.
710 (2021). Direct measurement of ATP7B peptides is highly effective in the diagnosis of Wilson
711 disease. *Gastroenterology*, *160*, 2367-2382.
- 712 8. Collins, A. G. (2018). The tortoise and the hare: Interactions between reinforcement learning
713 and working memory. *Journal of cognitive neuroscience*, *30*, 1422-1432.
- 714 9. Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working
715 memory, not reinforcement learning? A behavioral, computational, and neurogenetic
716 analysis. *European Journal of Neuroscience*, *35*, 1024-1035.
- 717 10. Collins, A. G., & Frank, M. J. (2018). Within-and across-trial dynamics of human EEG reveal
718 cooperative interplay between reinforcement learning and working memory. *Proceedings of
719 the National Academy of Sciences*, *115*, 2502-2507.
- 720 11. Collins, A. G., Ciullo, B., Frank, M. J., & Badre, D. (2017a). Working memory load
721 strengthens reward prediction errors. *Journal of Neuroscience*, *37*, 4332-4342.
- 722 12. Collins, A. G., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017b). Interactions
723 among working memory, reinforcement learning, and effort in value-based choice: A new
724 paradigm and selective deficits in schizophrenia. *Biological psychiatry*, *82*, 431-439.
- 725 13. Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working
726 memory contributions to reinforcement learning impairments in schizophrenia. *Journal of
727 Neuroscience*, *34*, 13747-13756.
- 728 14. Cremer, A., Kalbe, F., Gläscher, J., & Schwabe, L. (2021). Stress reduces both model-based
729 and model-free neural computations during flexible learning. *NeuroImage*, *229*, 117747.
- 730 15. Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-
731 trial EEG dynamics including independent component analysis. *Journal of neuroscience
732 methods*, *134*, 9-21.
- 733 16. Elzinga, B. M., & Roelofs, K. (2005). Cortisol-induced impairments of working memory
734 require acute sympathetic activation. *Behavioral neuroscience*, *119*, 98.
- 735 17. Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold your horses:
736 impulsivity, deep brain stimulation, and medication in parkinsonism. *Science*, *318*, 1309-1312.
- 737 18. Hamilton, D. A., & Brigman, J. L. (2015). Behavioral flexibility in rats and mice:
738 contributions of distinct frontocortical regions. *Genes, Brain and Behavior*, *14*, 4-21.

- 739 19. Geana, A., Barch, D. M., Gold, J. M., Carter, C. S., MacDonald III, A. W., Ragland, J. D.,
740 Silverstein S. M., & Frank, M. J. (2022). Using Computational Modeling to Capture
741 Schizophrenia-Specific Reinforcement Learning Differences and Their Implications on Patient
742 Classification. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 7, 1035-
743 1046.
- 744 20. Gold, J. M., Waltz, J. A., Matveeva, T. M., Kasanova, Z., Strauss, G. P., Herbener, E. S.,
745 Collins A.G.E., & Frank, M. J. (2012). Negative symptoms and the failure to represent the
746 expected reward value of actions: behavioral and computational modeling evidence. *Archives*
747 *of general psychiatry*, 69, 129-138.
- 748 21. Goldfarb EV, Froböse, MI, Cools R, & Phelps EA (2017). Stress and cognitive flexibility:
749 cortisol increases are associated with enhanced updating but impaired switching. *Journal of*
750 *Cognitive Neuroscience*, 29,14-24
- 751 22. Jafarpour, A., Buffalo, E. A., Knight, R. T., & Collins, A. G. (2022). Event segmentation
752 reveals working memory forgetting rate. *Isience*, 25, 103902.
- 753 23. Jaskir, A., & Frank, M. J. (2022). On the normative advantages of dopamine and striatal
754 opponency for learning and choice. bioRxiv 483879.
755 <https://doi.org/10.1101/2022.03.10.483879>.
- 756 24. Kim, J., Lee, H., Han, J., & Packard, M. (2001). Amygdala is critical for stress-induced
757 modulation of hippocampal long-term potentiation and learning. *Journal of neuroscience*, 21,
758 5222-5228.
- 759 25. Klein, T. A., Ullsperger, M., & Jocham, G. (2017). Learning relative values in the striatum
760 induces violations of normative decision making. *Nature communications*, 8, 1-12.
- 761 26. Leong, K. C., & Packard, M. G. (2014). Exposure to predator odor influences the relative use
762 of multiple memory systems: role of basolateral amygdala. *Neurobiology of Learning and*
763 *Memory*, 109, 56-61.
- 764 27. Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy
765 update rather than value prediction. *Journal of Neuroscience*, 31, 5504-5511.
- 766 28. Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis
767 of event-related potentials. *Frontiers in human neuroscience*, 8, 213.
- 768 29. Matuschek, H., Kliegl, R., Vasisht, S., Baayen, H., & Bates, D. (2017). Balancing Type I
769 error and power in linear mixed models. *Journal of Memory and Language*, 94, 305-315.
- 770 30. Meier, J. K., Staresina, B. P., & Schwabe, L. (2022). Stress diminishes outcome but enhances
771 response representations during instrumental learning. *Elife*, 11, e67517.
- 772 31. Oberauer, K., Farrell, S., Jarrold, C., and Lewandowsky, S. (2016). What limits working
773 memory capacity?. *Psychological bulletin*, 142, 758.
- 774 32. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory
775 capacity protects model-based learning from stress. *Proceedings of the National Academy of*
776 *Sciences*, 110, 20941-20946.
- 777 33. Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of
778 value signals in reward and punishment learning. *Nature communications*, 6, 1-14.
- 779 34. Plessow, F., Kiesel, A., & Kirschbaum, C. (2012). The stressed prefrontal cortex and goal-
780 directed behaviour: acute psychosocial stress impairs the flexible implementation of task
781 goals. *Experimental brain research*, 216, 397-408.
- 782 35. Quaedflieg, C. W. E. M., Stoffregen, H., Sebalo, I., & Smeets, T. (2019). Stress-induced
783 impairment in goal-directed instrumental behaviour is moderated by baseline working
784 memory. *Neurobiology of learning and memory*, 158, 42-49

- 785 36. R Core Team (2020). R: A language and environment for statistical computing. R Foundation
786 for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- 787 37. Rac-Lubashevsky, R., & Frank, M. J. (2021). Analogous computations in working memory
788 input, output and motor gating: Electrophysiological and computational modeling evidence.
789 *PLoS computational biology*, *17*, e1008971.
- 790 38. Raio, C. M., Hartley, C. A., Orederu, T. A., Li, J., & Phelps, E. A. (2017). Stress attenuates
791 the flexible updating of aversive value. *Proceedings of the National Academy of Sciences*,
792 *114*, 11241-11246.
- 793 39. Roozendaal, B., Okuda, S., De Quervain, D. F., & McGaugh, J. L. (2006). Glucocorticoids
794 interact with emotion-induced noradrenergic activation in influencing different memory
795 functions. *Neuroscience*, *138*, 901-910.
- 796 40. Schwabe, L., & Schächinger, H. (2018). Ten years of research with the Socially Evaluated
797 Cold Pressor Test: Data from the past and guidelines for the future.
798 *Psychoneuroendocrinology*, *92*, 155-161.
- 799 41. Schwabe, L., & Wolf, O. T. (2009). Stress prompts habit behavior in humans. *The Journal of*
800 *Neuroscience*, *29*, 7191-7198.
- 801 42. Schwabe, L., & Wolf, O. T. (2011). Stress-induced modulation of instrumental behavior: from
802 goal-directed to habitual control of action. *Behavioural brain research*, *219*, 321-328.
- 803 43. Schwabe, L., & Wolf, O. T. (2012). Stress modulates the engagement of multiple memory
804 systems in classification learning. *The Journal of Neuroscience*, *32*, 11042-11049.
- 805 44. Schwabe, L., Haddad, L., & Schachinger, H. (2008). HPA axis activation by a socially
806 evaluated cold-pressor test. *Psychoneuroendocrinology*, *33*, 890-895.
- 807 45. Schwabe, L., Höffken, O., Tegenthoff, M., & Wolf, O. T. (2011). Preventing the stress-
808 induced shift from goal-directed to habit action with a β -adrenergic antagonist. *Journal of*
809 *Neuroscience*, *31*, 17317-17325.
- 810 46. Simon-Kutscher, K., Wanke, N., Hiller, C., & Schwabe, L. (2019). Fear without context: acute
811 stress modulates the balance of cue-dependent and contextual fear learning. *Psychological*
812 *Science*, *30*, 1123-1135.
- 813 47. Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review.
814 *Neuroscience and Biobehavioral Reviews*, *36*, 1228-1248.
- 815 48. Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1994). Testtheoretische Analysen der
816 Mehrdimensionalen Befindlichkeitsfragebogens (MDBF) [Test-theoretical analyses of the
817 Multidimensional Mood State Questionnaire]. *Diagnostica*, *40*, 320-328.
- 818 49. Vaessen, T., Hernaes, D., Myin-Germeys, I., & van Amelsvoort, T. (2015). The dopaminergic
819 response to acute stress in health and psychopathology: a systematic review. *Neuroscience and*
820 *Biobehavioral Reviews*, *56*, 241-251.
- 821 50. Vogel, S., Fernández, G., Joëls, M., & Schwabe, L. (2016). Cognitive adaptation under stress:
822 a case for the mineralocorticoid receptor. *Trends in cognitive sciences*, *20*, 192-203.
- 823 51. Wimmer, G. E., & Poldrack, R. A. (2022). Reward learning and working memory: Effects of
824 massed versus spaced training and post-learning delay period. *Memory & cognition*, *50*, 312-
825 324.
- 826 52. Wirz, L., Bogdanov, M., & Schwabe, L. (2018). Habits under stress: mechanistic insights
827 across different types of learning. *Current Opinion in Behavioral Sciences*, *20*, 9-16.
- 828
- 829

830 **Figure 1.** Experimental protocol of the learning task and the two test phases. (A) In the learning phase,
831 in each block participants use deterministic reward feedback to learn which of three actions to select

832 for each stimulus image. The set size (or the number of stimuli; ns) varies from one to five across
 833 blocks. After each response feedback was presented audio-visually (see text for more detail). (B) The
 834 surprise reward-retention test protocol. In this task, participants are asked to recall the reward value of
 835 stimuli learned during the learning phase by choosing the stimulus they perceive to have been more
 836 rewarded within a pair of stimuli presented on every trial. (C) The surprise stimulus-response retention
 837 test protocol is a test of the learned stimulus-response “policy”. Here, participants are asked to recall
 838 the correct action for the probed stimulus. No feedback was given at either test phase.

839

840 **Figure 2.** Cooperative interaction between the RL and WM systems (adapted from Collins and Frank,
 841 2018): A. Both WM and RL inform expected Q values and thus inform reward prediction errors
 842 (RPEs). When the number of stimuli to learn (ssz or “set size”) is within WM capacity (e.g., ssz=2 on
 843 the left) the expected Q value of each contingency can be held in WM, thereby reducing RPE’s during
 844 early learning compared to those that would occur from RL alone. When set size exceeds WM
 845 capacity (e.g., ssz=5 on the right), degraded WM results in larger RPEs. B. Computational model
 846 simulations (recreated from Collins and Frank, 2018) capture the RL and WM interaction, showing
 847 that larger RPEs persist for longer when WM load is taxed (high ssz), thereby accumulating expected
 848 Q values in the RL system. C. Note that Q learning curves in panel B evolve more rapidly in high ssz,
 849 despite the opposite pattern in simulated behavioral learning curves (whereby WM contributes to rapid
 850 learning in low ssz.

851

852 **Figure 3.** Behavioral results from the learning phase. (A-B) Performance learning curves and reaction
 853 times (RT) for each set size as a function of the number of iterations of a stimulus (stim). (C)
 854 Performance as a function of WM load, the detrimental effect of delay is greater in high set sizes. (D-
 855 E) Reduced effects of both delay and set size as learning progresses from early (up to two previous
 856 correct choices) to late (the last two trials of each stimulus) trials in a block, suggestive of a transition
 857 from WM to RL.

858 **Figure 4.** Behavior performance at the test phase. (A) Effect of value difference and set size on the
 859 reward retention test performance. The proportion of correct selection of the more rewarding stimulus
 860 from a pair of the probed stimuli increases as a function of differences in the number of experienced
 861 rewards (Q value diff) and the set size in which they were learned. The median split of absolute value
 862 differences is shown (high-Q value difference trials depicted in red and low-Q value difference trials
 863 in blue). (B-C) Effect of set size on the stimulus-response retention test performance. The proportion
 864 of correct recall in the test phase increases as a function of the estimated Q values of the probed
 865 association and as a function of the set size in which it was learned. The median split of the estimated
 866 stimulus-response Q values is shown (high Q value associations in red and low Q value associations in
 867 blue). (D) Effect of EEG RL index on the reward retention test performance. The proportion of correct
 868 selection of the more rewarding stimulus from a pair of the probed stimuli increases as a function of
 869 the set size in which they were learned but was not further modulated by the magnitude of the EEG
 870 RL index of the stimuli. The median split of absolute differences in EEG RL indices is shown (high-
 871 EEG RL index difference in red and low-EEG RL index difference in blue). (E) Effect of the neural
 872 RL index on recall accuracy in the stimulus-response retention test. The neural RL index is shown as
 873 the median split across all the RL indices. Stimuli with high RL index are depicted in red and stimuli
 874 with low RL index are depicted in blue. (F) The EEG RL index increases parametrically with the
 875 increase in accumulated rewards. These neural learning curves parametrically increase with set size.
 876 Error bars represent standard errors.

877 **Figure 5.** EEG decoding of RL and WM effects during choice. Corrected event-related potentials
 878 (ERPs) exhibiting the effect of three main predictors (set-size in green, delay in blue, RL value
 879 quartiles in red; from top to bottom row) on the voltage of significant electrodes (FCz, CPz, and Poz
 880 for set size and delay, and FCz, CPz, and C3 for RL). The black line reflects the significant time points
 881 after permutation correction. On the right, the effect of each predictor in the row is exhibited with a
 882 scalp map topography at an early (300ms) and late (540ms) time points. The color in the scalp map
 883 represents significant thresholded t-values.

884

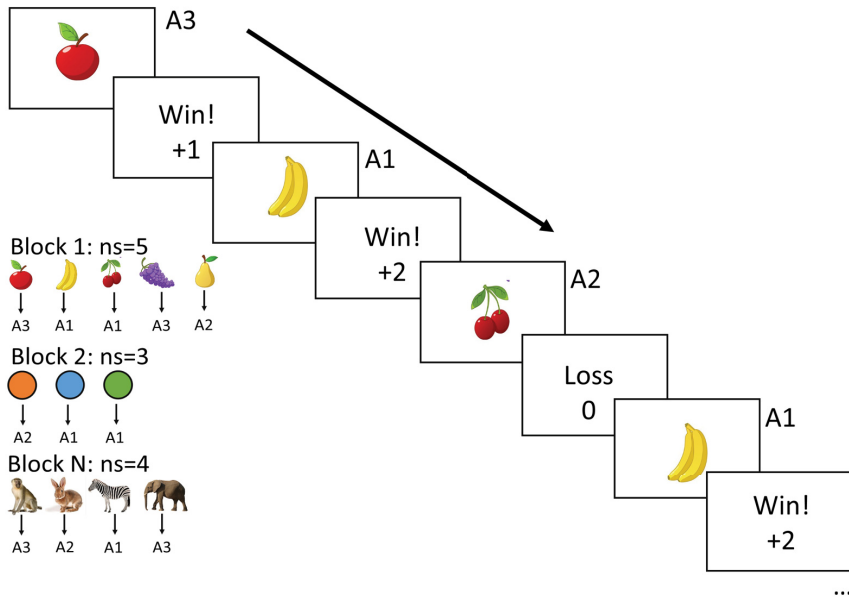
885 **Figure 6.** Successful stress induction. The exposure to the stressor led to significant increases in (A)
 886 systolic blood pressure, (B) diastolic blood pressure, (C) heart rate, and (D) salivary cortisol levels;
 887 error bars represent standard errors. The control group is depicted in dark blue and the stress group in
 888 red. ** $p < 0.01$, *** $p < 0.001$ for the comparison between the stress group and the control group.

889 **Figure 7.** Stress effects during the learning and test phases. (A) Learning curves across iterations as a
 890 function of set size in the control group (B) and stress group. (C) Learning curves across the number
 891 of previous correct as a function of delay (1 to 5 where 5 reflects delay of five and above) in the
 892 control group (D) and stress group. (E) Effect of stress on the reward retention test performance. The
 893 proportion of correct selection of the more rewarding stimulus from a pair of the probed stimuli
 894 increases as a function of the set size in both the control group (depicted in black) and in the stress
 895 group (depicted in red). (F) Effect of stress on recall accuracy in the stimulus-response retention test.
 896 The proportion of correct recall in the stimulus-response test increases as a function of the set size in
 897 both the control group (depicted in black) and the stress group (depicted in red). Error bars represent
 898 standard errors.

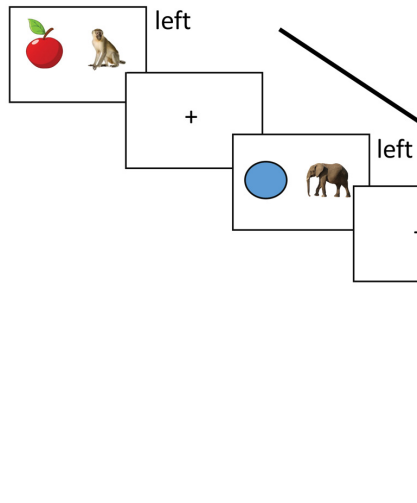
899 **Table 1.** Subjective mood and procedure ratings across the experiment in both control and stress
 900 groups. The mean and standard deviation of the ratings before and after the procedures are reported for
 901 the control group (upper part) and for the stress group (bottom part).

	Control group		
	Before	After	End of testing day
Subjective mood			
Depressed mood vs. elevated mood	33.69 (4.99)	34.26 (4.72)	33.86 (4.66)
Restlessness vs. calmness	32.476 (6.08)	33.83 (5.14)	33.24 (4.61)
Sleepiness vs. wakefulness	28.571 (6.48)	28.31 (6.88)	26.64 (6.78)
Rating of control procedure			
difficult	-	4.09 (13.21)	-
unpleasant	-	9.52 (21.88)	-
stressful	-	4.20 (15.23)	-
painful	-	3.79 (14.62)	-
	Stress group		
	Before	After	End of testing day
Subjective mood			
Depressed mood vs. elevated mood	33.76 (3.51)	31.57 (5.32)	33.43 (3.99)
Restlessness vs. calmness	32.99 (4.24)	30.45 (6.14)	32.43 (4.72)
Sleepiness vs. wakefulness	28.98 (5.71)	29.86 (6.16)	26.45 (6.12)
Rating of stressor			
difficult	-	50.69 (28.01)	-
unpleasant	-	58.73 (28.09)	-
stressful	-	40.17 (26.70)	-
painful	-	55.40 (25.97)	-

A Learning task



B Reward retention test



C Stimulus-response retention test

