Interactions Between Working Memory, Reinforcement Learning and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia

Supplementary Information

Computational Model

We describe here the model simulated in Figure 1. This model is a version of the model published in (1), adapted to the changes in this protocol.

RLWM Model

RLWM includes two modules, a classic incremental RL module with learning rate α, and a WM module that can learn in a single trial (learning rate 1) but is capacity-limited (with capacity K). The WM module is also subject to forgetting. The final action choice is controlled by weighing the contributions of the RL and WM modules' policies. How much weight is given to WM relative to RL (the mixture parameter) depends on two factors. First, it depends on what the probability is that a stimulus is stored in WM of capacity K. If there are fewer stimuli than WM can hold (n_S<=K), then that probability is 1. Otherwise, only K out of ns can be stored. Second, the overall reliance of WM vs. RL is scaled by factor 0<ρ<1, with higher values reflecting relative greater confidence in WM function. Thus, the weight given to the WM policy relative to RL policy is w = ρ x min(1, K/n_S).

Reinforcement learning module: This is a standard RL module with simple delta rule learning. For each stimulus s, and action a, the expected reward Q(s,a) is learned as a function of reinforcement history. Specifically, the Q value for the selected action given the stimulus is

updated upon observing each trial's reward outcome r_t (1 or 2 for correct, 0 for incorrect) as a function of the prediction error between expected and observed reward at trial *t*:

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

where $\delta_t = r_t - Q_t(s,a)$ is the prediction error, and α is the learning rate. Choices are generated probabilistically with greater likelihood of selecting actions that have higher Q values, using the softmax choice rule:

$\pi_{RL}(a|s) = \exp(\beta Q(s,a)) / \Sigma_i(\exp(\beta Q(s,a_i)))$.

Here, β is an inverse temperature determining the degree with which differences in Q- values are translated into more deterministic choice, and the sum is over the three possible actions a_i.

Working memory: To implement an approximation of a rapid updating but capacity-limited WM, we model a module that learns stimulus action values Q_{WM} similarly to the RL module, but with three differences: 1) learning rate $\alpha_{WM} = 1$ to represent fast learning, 2) outcome is 1 for correct, 0 for incorrect (rather than the observed reward), and 3) WM values are subject to decay. Furthermore, to mode capacity limitation such that only at most K stimuli can be remembered, we assume that at any time, the probability of a given stimulus being in working memory is $p_{WM} = \rho \times \min(1, K/n_S)$. Thus, the overall policy is:

$\pi = p_{WM}\pi_{WM} + (1-p_{WM})\,\pi_{RL}$.

where π_{WM} is the WM softmax policy, and π_{RL} is the RL module's policy.

Forgetting: We allow for potential decay or forgetting in Q_{WM}-values on each trial, additionally updating all Q_{WM}-values at each trial, according to:

$Q_{WM} \leftarrow Q_{WM} + \phi\,(Q_0 - Q_{WM})$,

where $0<\phi<1$ is a decay parameter pulling at each trial the estimates of values towards initial value $Q_0 = 1/n_A$.

Perseveration: To allow for potential neglect of negative, as opposed to positive feedback, we estimate a perseveration parameter *pers* such that for negative prediction errors (delta<0), the learning rate α is reduced by α = (1-*pers*) x α. Thus values of *pers* near 1 indicate perseveration with complete neglect of negative feedback, whereas values near 0 indicate equal learning from negative and positive feedback. This is applied to both RL and WM module.

Undirected noise: The softmax temperature allows for stochasticity in choice in an oriented way, by making more valuable actions more likely. We also allow for "slips" of action ("irreducible noise", i.e., even when Q value differences are large). Given a model's policy π = p(a|s), adding undirected noise consists in defining the new mixture policy:

$\pi' = (1-\varepsilon)\,\pi + \varepsilon U,$

where U is the uniform random policy (U(a) = $1/n_A$, n_A=3), and the parameter $0<\varepsilon<1$ controls the amount of noise (2–4).

Figure 1E results came from 100 Simulations of the computational model with the new design, for two sets of parameters sharing α=0.1, β=8, ϕ=0.1, ε=0.05), and either K=2 and ρ=0.8 (poor WM use) or K=3 and ρ=0.9 (good WM use).

Specific Experimental Methods

Experiment 1: This experiment included 22 blocks (3 blocks of *ns*=1, 4, 5 and 6 each, 6 blocks of *ns*=3, and 4 blocks of set size *ns*=2). Subjects encountered 15 iterations of each stimulus per block, pseudo-randomly interleaved. Participants had 1.4 s to respond, and observe feedback for 0.5 s, followed by 0.5 - 0.8 s fixation cross before the start of the next trial. Twenty-nine healthy Brown University undergraduates (11 men, 18 women, age 18-29, mean 21.7)

participated in this task; one was excluded from data analysis because of a technical issue during testing. The task lasted ~ 1h, and included 213 pairs in test phase. Pairs of images in each trial were pseudo-randomly selected to sample across all possible pairs, ensuring good representation of all set size pairs, block pairs and probability pairs. All 75 different stimuli encountered in the learning phase were presented in the test phase at least once.

Experiment 2: The trial timing was identical to experiment 1. Experiment 2 included 12 blocks (4 of set size 2, 3 of set size 3, 2 of set size 4, 2 of set size 5). Subjects encountered 13 iterations of each stimulus per block, pseudo-randomly interleaved. Pairs of images in each trial were pseudo-randomly selected to sample across all possible pairs, ensuring good representation of all set size pairs, block pairs and probability pairs. There were 156 pairs in the test phase, and all 39 different stimuli encountered in the learning phase were presented in the test phase at least once. The full experiment lasted ~ ½ h. Half of these participants performed another, unrelated learning task first. 52 young healthy participants (25 men, 27 women, age 18-27, mean 19.6) participated in the experiment. One participant was excluded from data analysis for a technical problem during testing. The results were identical if we restricted analysis to participants who performed this experiment first.

Experiment 3: PSZ-HC study: The design was identical to experiment 2, with only timing differences: participants had 3 s to answer, feedback was presented for 0.6-1s, and the ITI was 0.8 – 1.2 s; EEG was measured during participants' performance. 49 patients, 32 matched controls performed the task. Two patients were excluded for performance indicating lack of engagement with the task (P(correct)<0.5 over the whole learning phase).

Forty-nine participants with a diagnosis of schizophrenia or schizoaffective disorder (according to DSM-IV diagnostic criteria) and 32 controls were recruited for the experiment. Patients were

clinically and pharmacologically (drug and dose) stable (> 4 weeks) outpatients from the Maryland Psychiatric Research Center or other nearby clinics. Controls were free from a lifetime history of SZ, other psychotic disorder, current Axis I disorder, drug dependence, neurological disorder, or cognitively impairing medical disorder, with no family history of psychosis in first-degree relatives. Controls were screened with the Structured Clinical Interview for DSM-IV.

Analysis details

Learning phase: Missed trials or trials with reaction times <200 ms were excluded (Exp. 1: average 2 trials, max 8; Exp. 2: average 1 trial, max 4.2; patients/controls: average 0.7, max 7). We analyzed the proportion of correct choices as a function of the variables: *set size* (number of stimulus images in the block), *iteration* (how many times the stimulus has been encountered), *pcor* (number of previous correct choices for the current stimulus), and *delay* (number of trials since the last correct choice for the current trial's stimulus). Learning curves (Fig. 2A-B) were obtained by taking the proportion of correct responses/average reaction times as a function of set size and iteration. We also investigated performance as a function of set size (high: ns>=4, low: ns<=3), and $p(r=2|correct)$ (Fig. 2D). Effects of delay were visualized by averaging the proportion of correct trials as a function of delay and set size (Fig. 2E), or delay and pcor (Fig. 2C). We label "early" trials as trials with 1 or 2 previous correct, and late trials as trials with n or n-1 previous correct trials for this specific stimulus (with n being the maximum number of previous correct trials).

Train logistic regression: To quantify the effect of working memory and RL on a trial-by-trial basis, we modeled each participant's choices using logistic regression. Specifically, each trial's probability of a correct choice was modelled as a function of *set size*, *pcor*, and *delay* (we excluded set size one for this analysis because of the lack of variability in *delay)*. We

transformed each predictor by $X \diamond -1/X$ because we observed in previous experiments that this leads to better fits (e.g., for set size ns, there are larger performance deficits between ns=4 and 3 than between ns=6 and 5, and this is captured by using 1/ns as a predictor). Results are similar without this transformation. We first investigated a model with only main effects, then include the interaction between the three factors.

Test logistic regression: To analyze choices in the test phase data, we defined for each image the following characteristics: value (reward history: **average of all feedback received for this image**), set size and block (the set size and block number of the block in which the stimulus image was encountered).

To visualize test performance, we separated test trials into three bins as a function of the absolute reward value difference between the two items. Performance in each bin was defined as the proportion of times the image with the higher value was chosen (Fig. 5 left).

To analyze test performance we used a logistic regression to model the choice of the right vs. left image, allowing us to assess various factors that could modulate the effective preferences. Specifically, we consider the following predictors:

ΔQ = *value(right)-value(left)*, assessing value difference effects.

Δns = *ns(right)-ns(left)*, assessing whether subjects prefer items that had been encountered in high or low set-sizes independently of experienced value, as might be expected if the experience of cognitive effort in high set sizes is aversive.

Δblock = *block(right)-block(left),* assessing whether there is an effect of recency as might be expected if values decay with time until test phase.

stay = *1* if *choice(t-1)*=right, *-1* otherwise, assessing response autocorrelation in the test phase.

To investigate whether the effect of value is modulated by other factors, we consider two additional predictors:

Mean(Q)*ΔQ: assesses whether there is a bias in choice discrimination, i.e. whether the ΔQ value effect is stronger or weaker for choices among items with relatively high or low mean Q values. Previous studies have indicated that manipulations that increase striatal dopamine improve choice discriminations in test phase among items with high mean Q value, whereas manipulations that decrease striatal dopamine improve relative value discrimination among items with low mean Q value (5–7).

Mean(*ns)*ΔQ: assesses whether value discrimination is stronger or weaker when the items came from relatively high or low set sizes.

Finally, we investigated two methods to define the value of an image for the logistic regression: 1) the simple empirical average number of points received for this image; 2) a weighted average of the points received allowing us to assess via two free parameters i) the relative subjective value of 1 vs. 2 points and ii) the contribution of correct (irrespective of reward magnitude) vs. incorrect trials to the value of the image.

Controlling for model complexity with AIC, we find that a model including all those predictors fits better than a model including a subset of them, or not parameterizing value (though the latter leads to similar results). We thus report results from this full model. We only include participants for which the model fits better than chance in the analysis of the regression weights; this includes 24 out of 27 participants in exp. 1, 43 out of 51 participants in exp. 2, 29 out of 32 healthy controls, and 42 out of 48 people with SZ.

Supplementary Results

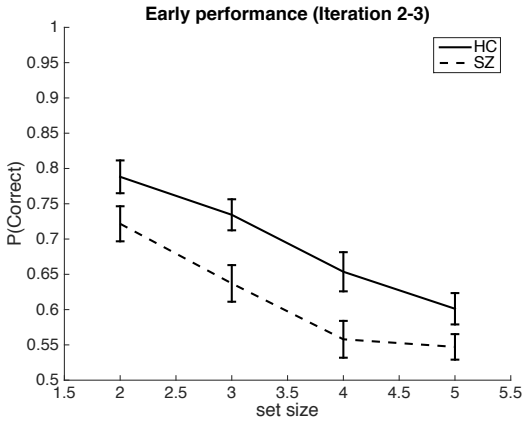

Figure S1. Early learning performance: A deficit in working memory predicts that patients should show a deficit in early learning trials compared to healthy controls. We define early learning as iteration 2 and 3 of each image (corresponding to the first trial with previous information and the last potential trial without information, if assuming perfect memory). We find a strong deficit (t(77)=3.08, p=.003).

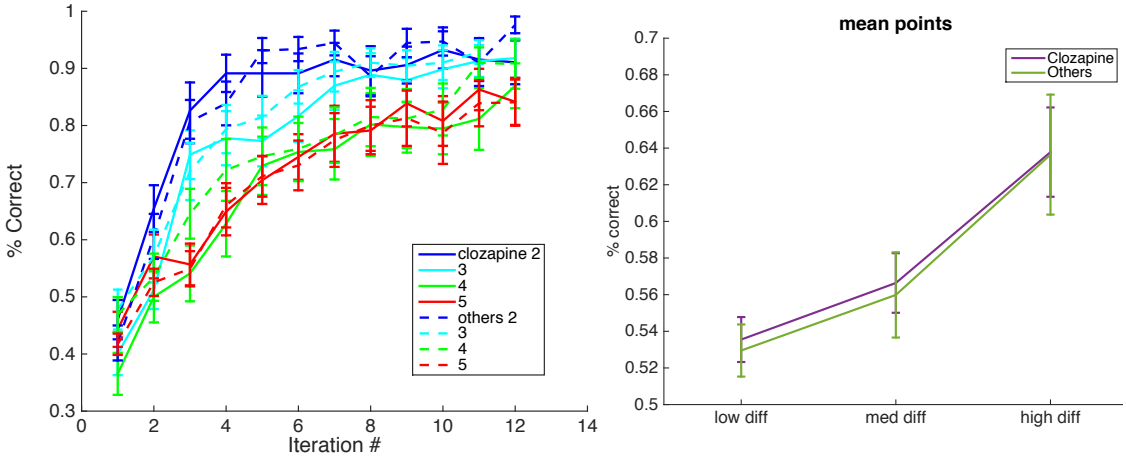

Figure S2. Effect of medication type in patients: Learning phase (left) and test phase (right) results plotted for 23 patients under Clozapine and 21 patients under different medication. There were no differences.

Supplementary Discussion

Relationship to the model-based/model-free RL framework

Others have proposed a dichotomy of systems contributing to learning by reinforcement. In particular, the framework of model-free vs. model-based behavior has recently gained lots of traction, in particular due to the development of a task allowing to simultaneously track their contribution (2-step task, (8)). One may ask how our dichotomy of RL vs. WM compares to this framework. There are similarities; in particular, model-based learning most likely relies on executive functions, and in particular on working memory, as indicated by the loss of model-based contributions to behavior under dual task circumstances (9). However, we think our dissociation is different in important ways.

Note that indicators of model-based learning are usually in this literature limited to environments that include sequential choices affording the potential to plan ahead. In our case, there is no sequential dependence between iterative choices and thus no reason for planning or model-based behavior. Indeed, within a given set size, this type of learning would typically be modeled with purely model-free RL. While it is possible to frame the WM aspect of our model as model-based (since it involves the predicted state-action-outcome association for each stimulus), our dual-systems dissociation shows a role of working memory even in learning behavior that would typically be considered model-free, and so is in many ways orthogonal to it.

In particular, previous findings showed that model-free choice in sequential behavior did not appear to require WM resources (9). Our findings contribute to the literature on model-based/model-free learning by showing that this is not necessarily generalizable to all *apparently* model-free behavior: indeed, behavior in our task is apparently model-free as defined by that literature, as it does not require forward planning, but imposes considerable WM requirements.

Clinical symptoms and medication

An important issue is the role of medication in our findings. There is good evidence in the literature that chronic, medicated PSZ have different learning deficits, when compared with unmedicated patients; for example, a recent study (10) showed deficits specifically in negative-RPE-driven learning in unmedicated patients. Addressing the impact of antipsychotic drugs (APDs) on learning and decision making in medicated patients is a difficult issue, given that psychosis has been associated with elevated synaptic dopamine (11,12) and all effective APDs are thought to modulate dopamine D2 receptors (13). The end functional result of this combination on striatal function and its manifestation as a neurocognitive profile is difficult to ascertain (14). One hypothesis is that antipsychotics normalize the striatal/dopaminergic associated trial-by-trial aspects of learning (and specifically the balance of positive and negative RPEs (15,16)) more than influencing the PFC associated working-memory processes. Under this hypothesis, our study is in accordance with (16), which also concluded that statistical learning about RPEs was intact in medicated patients, but that there were deficits in expected value computation associated with PFC function, similar to the WM deficit observed here. However, a simple normalization of striatal function during RL by antipsychotic administration in schizophrenia is likely too simplistic, as it is unlikely that chronic dopamine blockade would preserve the integrity of high fidelity learning signals (14). We simply note here that our results were not linked to medication dosage, and that patients on clozapine had similar behavior to non-clozapine patients (see supplement), but future research is needed to better understand this issue.

Furthermore, our results did not provide insight as to whether specific symptoms (beyond cognitive symptoms), in particular negative symptoms, were linked to distinct contributions to learning. The fact that we do not see a negative symptom signal here is interesting from the standpoint that we have observed negative symptom correlations with other PFC-related

processes, but not all. PFC-related processes that have been found to be linked to negative symptoms include win-stay/lose-shift, uncertainty-driven exploration (17), willingness to expend effort (18,19), and the extent to which RL relies on a Q-learning mechanisms (16). With multiple other measures of learning and motivation, we have observed correlations with standard neurocognitive scores, but not with negative symptoms scores. These include probabilistic reversal learning (20,21); on-the-fly expected value computation (22); Iowa Gambling Test performance (23); susceptibility to confirmation bias (24); and distractor devaluation effects (25). It is possible that the PFC-dependent working memory process that influences learning from reinforcement, as observed in our study, is more similar to the latter kinds of behavior than the former. This would make our finding coherent with a broader literature.

Supplemental References

1. Collins a. GE, Brown JK, Gold JM, Waltz J a., Frank MJ. Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. J Neurosci. 2014 Oct 8;34(41):13747–56.
2. Collins A, Koechlin E. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. O'Doherty JP, editor. PLoS Biol. 2012 Mar 27;10(3):e1001293.
3. Collins AGE, Frank MJ. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychol Rev. 2013;120(1):190–229.
4. Guitart-Masip M, Huys QJM, Fuentemilla L, Dayan P, Duzel E, Dolan RJ. Go and no-go learning in reward and punishment: interactions between affect and effect. Neuroimage. 2012 Aug 1;62(1):154–66.
5. Collins AGE, Frank MJ. Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive.
6. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004 Dec 10;306(5703):1940–3.
7. Frank MJ, Santamaria A, Reilly RCO, Willcutt E. Testing Computational Models of Dopamine and Noradrenaline Dysfunction in Attention Deficit / Hyperactivity Disorder. 2007;1583–99.
8. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011 Mar 24;69(6):1204–15.
9. Otto AR, Gershman SJ, Markman AB, Daw ND. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psychol Sci. 2013 May 4;24(5):751–61.
10. Reinen JM, Van Snellenberg JX, Horga G, Abi-Dargham A, Daw ND, Shohamy D. Motivational Context Modulates Prediction Error Response in Schizophrenia. Schizophr Bull. 2016;42(6):1467–75.
11. Howes OD, Kambeitz J, Kim E, Stahl D, Slifstein M, Abi-Dargham A, et al. The nature of dopamine dysfunction in schizophrenia and what this means for treatment. Arch Gen Psychiatry. 2012 Aug;69(8):776–86.
12. Laruelle M, Abi-Dargham A. Dopamine as the wind of the psychotic fire: new evidence from brain imaging studies. J Psychopharmacol. SAGE PublicationsLondon, Thousand Oaks, CA and New Delhi; 1999 Jul;13(4):358–71.
13. Kapur S, Zipursky RB, Remington G. Clinical and theoretical implications of 5-HT2 and D2 receptor occupancy of clozapine, risperidone, and olanzapine in schizophrenia. Am J Psychiatry. 1999 Feb;156(2):286–93.
14. Maia T V., Frank MJ. An Integrative Perspective on the Role of Dopamine in Schizophrenia. Biol Psychiatry. Elsevier; 2017;81(1):52–66.
15. Insel C, Reinen J, Weber J, Wager TD, Jarskog LF, Shohamy D, et al. Antipsychotic dose modulates behavioral and neural responses to feedback during reinforcement learning in schizophrenia. Cogn Affect Behav Neurosci. Springer US; 2014 Mar 21;14(1):189–201.
16. Gold JM, Waltz JA, Matveeva TM, Kasanova Z, Strauss GP, Herbener ES, et al. Negative symptoms and the failure to represent the expected reward value of actions:

Behavioral and computational modeling evidence. Arch Gen Psychiatry. 2012;69(2).

17. Strauss GP, Robinson BM, Waltz J a, Frank MJ, Kasanova Z, Herbener ES, et al. Patients with schizophrenia demonstrate inconsistent preference judgments for affective and nonaffective stimuli. Schizophr Bull. 2011 Nov;37(6):1295–304.

18. Gold JM, Strauss GP, Waltz J a, Robinson BM, Brown JK, Frank MJ. Negative Symptoms of Schizophrenia Are Associated with Abnormal Effort-Cost Computations. Biol Psychiatry. Elsevier; 2013 Feb 7;1–7.

19. Barch DM, Treadway MT, Schoen N. Effort, anhedonia, and function in schizophrenia: Reduced effort allocation predicts amotivation and functional impairment. J Abnorm Psychol. American Psychological Association; 2014;123(2):387–97.

20. Waltz J a, Gold JM. Probabilistic reversal learning impairments in schizophrenia: further evidence of orbitofrontal dysfunction. Schizophr Res. NIH Public Access; 2007 Jul 1;93(1–3):296–303.

21. Reddy LF, Waltz JA, Green MF, Wynn JK, Horan WP. Probabilistic Reversal Learning in Schizophrenia: Stability of Deficits and Potential Causal Mechanisms. Schizophr Bull. Oxford University Press; 2016 Jul;42(4):942–51.

22. Brown JK, Waltz JA, Strauss GP, McMahon RP, Frank MJ, Gold JM. Hypothetical decision making in schizophrenia: The role of expected value computation and “irrational” biases. Psychiatry Res. 2013;209(2):142–9.

23. Brown EC, Hack SM, Gold JM, Carpenter WT, Fischer BA, Prentice KP, et al. Integrating frequency and magnitude information in decision-making in schizophrenia: An account of patient performance on the Iowa Gambling Task. J Psychiatr Res. 2015;66:16–23.

24. Doll BB, Waltz J a, Cockburn J, Brown JK, Frank MJ, Gold JM. Reduced susceptibility to confirmation bias in schizophrenia. Cogn Affect Behav Neurosci. 2014 Jan 31;

25. Strauss GP, Lee BG, Waltz JA, Robinson BM, Brown JK, Gold JM. Cognition-emotion interactions are modulated by working memory capacity in individuals with schizophrenia. Schizophr Res. 2012;141(2):257–61.