Supporting Text

Methods

Procedures were approved by the University of Colorado Human Research Committee. Participants sit in front of a computer screen in a lighted room and view pairs of visual stimuli that are not easily verbalized (Japanese Hiragana characters). These stimuli are presented in black on a white background, in 72 pt font. Instructions given to participants are as follows:

Two figures will appear simultaneously on the computer screen. One figure will be correct and the other will be incorrect, but at first you won't know which is which! There is no ABSOLUTE right answer, but some symbols will have a higher chance of being correct. Try to pick the symbol that you find to have the highest chance of being correct. Press the key labeled L to select the symbol on the left, and the key labeled R to select the symbol on the right. At first you may be confused, but don't worry, you'll have plenty of practice!

After each choice, visual feedback is immediately provided (duration 1.5 seconds), consisting of the word "Correct!" printed in blue or "Incorrect" printed in red. If no response is made within four seconds, the words "no response detected" are printed in red.

In the training phase, three different stimulus pairs (AB, CD, EF) are presented in random order, with the assignment of Hiragana character to stimulus elements A-F counterbalanced across subjects. Feedback follows the choice to indicate whether it was correct or incorrect, but this feedback is probabilistic. Choosing stimulus A leads to correct (positive) feedback in 80% of AB trials, whereas choosing stimulus B leads to incorrect (negative) feedback in these trials. CD and EF pairs are less reliable: stimulus C is correct in 70% of CD trials, while E is correct in 60% of EF trials. Over the course of training participants learn to choose stimuli A, C and E more often than B, D, or F.

We enforced a performance criterion (evaluated after each training block of 60 trials) to ensure that all participants were at the same performance level before advancing to test. Because of the different probabilistic structure of each stimulus pair, we used a different criterion for each (65% A in AB, 60% C in CD, 50% E in EF). (In the EF pair, stimulus E is correct 60% of the time, but this

is particularly difficult to learn. We therefore used a 50% criterion for this pair simply to ensure that if participants happened to "like" stimulus F at the outset, they nevertheless had to learn that this bias was not going to consistently work.). The participant advanced to the test session if all these criteria were met, or after six blocks (360 trials) of training.

Participants were subsequently tested with the same training pairs, in addition to all novel combinations of stimuli, in random sequence. Prior to the test phase they were given the following instructions: *"It's time to test what you've learned! During this set of trials you will NOT receive feedback ('Correct' or 'Incorrect' to your responses. If you see new combinations of symbols in the test, please choose the symbol that 'feels' more correct based on what you learnt during the training sessions. If you're not sure which one to pick, just go with your gut instinct!"*

Each test pair was presented four times for a maximum of four seconds duration, and no feedback was provided.

Additional Data and Analysis

DRD2 Gene and RT Effects.

In the main paper we showed reaction time slowing effects for *DRD2* genes and avoid-B performance. Overall, RT's were slower for avoid-B trials than choose-A trials ($F[1,66] = 7.4$, $P = 0.008$), as expected since it takes longer to avoid a stimulus (B) in order to select the alternative than it does to just directly choose one (A). Notably, the degree of slowing was reduced with increasing number of T alleles ($F[1,66] = 3.6$, $P = 0.06$), again showing enhanced NoGo function. This effect of RT slowing was not observed for the *DARPP-32* gene ($F[1,66] = 0.8$). Finally, these selective *DRD2* genetic effects on NoGo learning were found despite no effect on overall reaction times to positive or negative test trials alone (all P's > 0.25).

COMT Gene and Trial-to-Trial Switching and RT Effects.

As reported in the main paper, *COMT* effects on switching were significant even in the very first five training trials of each type. Here we show that this was due to met allele carriers starting out with high degree of switching and slowing following negative feedback, but reducing this behavior as training progressed (SI Fig. 6). This is consistent with the notion that participants begin with a "prefrontal strategy" and are therefore sensitive to working memory for recent trial outcomes, but then this trial-to-trial information becomes less informative as they integrate probabilistic feedback

over many trials. This is also consistent with observations that PFC is preferentially active during new learning and less so as cue-response associations become more familiar (*1,2*).

To analyze the effects of *COMT* on switching and slowing as a function of training, we grouped the initial block of training into sub-blocks of five trials of each type (AB, CD, EF; 15 trials total). We then analyzed whether the effects of *COMT* on switching interacted with this sub-block variable (1,2,3,4). We found that switching interacted with sub-block, such that met allele carriers started out with increased switching and slowing following single negative feedback experiences, which decreased across sub-blocks (SI Fig. 6). This is consistent with the notion that these subjects begin with a "prefrontal strategy", and are initially very influenced by the most recent reinforcement experience, but that this then gives way to systems that integrate reinforcement values over multiple trials.

There was no main effect of *COMT* genotype on overall reaction times ($F[1,66] = 0.05$). Nevertheless, the number of met alleles predicted the degree of post-error slowing when faced with decisions that were most recently associated with negative feedback ($F[1,66] = 4.2$, $P = 0.04$), and this slowing decreased with time (RT slowing interaction with training block; $F[1,66] = 6.1$, $P = 0.016$). This slowing effect is commonly observed as participants become more cautious during subsequent decisions (*3*). But just as in the switching findings, the post-error slowing observed here was specific to a particular stimulus context (ie it was not observed immediately following an error, but only when the relevant stimulus next appeared), and therefore depends on working memory for stimulus-reinforcement associations. Again, these sequential RT effects were not modulated by *DRD2* ($F[1,67] = 1.99$, ns) or *DARPP-32* genotypes ($F[1,67] = 0.6$).

Table 1 shows raw RT scores for each condition by genotype. The only significant effects were on RT differences in avoid-B - choose-A trials (for *DRD2*) and in trials following errors relative to following correct choices (for *COMT*).

Statistical Analysis

We performed a general linear model (GLM) regression to test the hypotheses in the main paper, using between subjects continuous factors (e.g. number of met alleles for *COMT* analysis or number of T alleles for *DRD2* analysis). Where appropriate, we also included repeated measures multivariate analyses to test for interaction effects (e.g. on choose-A vs avoid-B accuracy or reaction times). For choose-A and avoid-B accuracy statistics, percent choices were arcsine-

transformed (*4*) due to a moderate number of cases where values were 100% (similar results were obtained with non-transformed data).

Degrees of Freedom.

The number of degrees of freedom was not always the same in our analyses in the main paper. This was due to two simple factors. First, we were unable to obtain *COMT* genotypes for one subject, so the DF for all *COMT* analyses is one less than those for the other genes. Similarly, the DF was reduced by one for the within-subject reaction time analyses, comparing choose-A to avoid-B RTs as a function of *DRD2* genotype. This is because reaction time analyses were computed on correct trials. There was one subject who never responded correctly on avoid-B conditions, and therefore we did not have a RT measure for this subject.

Physiological Issues and Neural Model Considerations

We reported that increased *COMT* met allele expression, associated with elevated prefrontal DA levels, was predictive of trial-to-trial learning from negative outcomes. In contrast, we suggested that the striatal D2 receptor is necessary for learning from decreases in DA during negative feedback (and integrating this over multiple trials). How would elevated DA levels in PFC facilitate learning from negative events while decreases in DA support this learning in striatum? First, several studies suggest that prefrontal and striatal DA levels are inversely related (*5*–*7*). Second, pauses in DA firing during negative events occur only transiently (pause durations are roughly 200 ms). In striatum, reuptake is fast [4-6 μmols/ sec; (*8*)] and as a result, the half-life of DA in the synapse is short enough so that DA levels can sufficiently decrease during DA pauses to enhance striatal NoGo learning (*12*). In contrast, due to the lack of dopamine transporters in PFC, dopamine clearance in PFC is far slower [0.05 micromolar /sec; (*9*)] – it is therefore somewhat more unlikely that a transient pause in midbrain DA firing would have any effect on PFC DA concentration. Furthermore, available evidence shows that prefrontal DA levels actually increase in response to negative events over temporally extended periods (*10, 11*). In contrast, DA levels either generally do not change or actually decrease over this same time period in striatum (*10, 11*).

Note that in our neural model of prefrontal/striatal interactions, in addition to driving striatal NoGo learning, DA decreases (dips) also played a relatively minor role in enhancing negative relative to positive outcome representations in PFC (*12*). This aspect of our neural model differs

from the notion depicted here that *elevations* in PFC DA support rapid trial-to-trial adjustments due to negative outcomes. This disconnection between our neural model and the data (both our genetic data and physiological evidence quoted above), will force us to modify this role of DA in modulating PFC representations in future endeavors. In particular, we will follow along the lines of other prominent models of PFC DA (cited in the main paper) which demonstrate that DA enhances robust maintenance properties in the PFC; in our framework this can support working memory for recent reinforcement outcomes. This underscores the iterative nature of neural modeling, whereby models should be continually updated and improved in the face of challenging data. Nevertheless, we emphasize that the key prediction of our neural model, supported by the data, is that enhanced PFC function (often associated with *COMT* met allele expressions) should predict rapid trial-to-trial adaptation of behavior, but not slow integrative (BG-dependent) learning.

Q-learning: Methods, Justification, and Additional Analysis

Q learning (*13, 14*) is a mathematical model that simulates reinforcement-based decision making, and is able to fit participants trial-by-trial sequence of responses. Here, we apply the Q learning algorithm to the probabilistic selection task (*15*). The rationale for doing so is to disintegrate subjects' performance in this task into different components, as motivated by neurobiological models, and also to determine whether individual differences in model parameters are accounted for by genetic measures of interest. We implemented two parallel versions of the Q learning algorithm, maximizing fit to participants performance in either the training phase (which is standard for reinforcement tasks) or to subsequent generalization performance in the test phase (novel).

The Gain-Loss Q Learning Model.

Because of computational and experimental evidence suggesting that positive and negative reinforcement learning are subserved by disparate striatal mechanisms, our model incorporates two learning rate parameters, associated with loss and gain [negative and positive feedback; (*16*)]. Q learning models assume that subjects maintain independent estimates (Q values) of the reward expected for each stimulus. The expected value of selecting a stimulus i (where i can be A,B,C,D,E or F) is computed as follows:

$$Q_i(t+1) = Q_i(t) + \alpha_G[r(t) - Q_i(t)]_+ + \alpha_L[r(t) - Q_i(t)]_-, \quad \textbf{[5]}$$

where t is trial number, and all Q_i are initialized to 0. The best fitting learning rate parameters α_G and α_L to each participant's sequence of responses reflects the degree to which previous reinforcement outcomes affect subsequent Q values. Therefore a large learning rate is associated with a recency effect whereas a small learning rate suggests that Q values are being integrated over multiple trials. This analysis applies to both positive (α_G) and negative (α_L) outcome learning.

The probability of selecting one stimulus over another (eg. A over B) was computed as:

$$P_A(t) = \frac{e^{\frac{Q_A(t)}{\beta}}}{e^{\frac{Q_A(t)}{\beta}} + e^{\frac{Q_B(t)}{\beta}}}, \quad \textbf{[6]}$$

where β is an inverse gain parameter and reflects the participant's tendency to exploit (ie., to choose the stimulus with the currently highest Q value) or explore (eg., to randomly choose a response) (*17*). The same equation applies for other trial-types, replacing A and B with C,D,E,F as appropriate.

Fit to Training.

This model was first fit to each participant's training data, by searching through the space of each of three parameters, from 0.01 to 1 with a step size of 0.03. We then optimized the log likelihood estimate (LLE) fit of the model to each subjects behavioral choices:

$$LLE = log(\prod_t P_{i^*,t}), \quad \textbf{[7]}$$

where t is trial number and i^*, t denotes the subjects choice on trial t. For each subject, the best fit parameters are those associated with the maximum LLE value and are, by definition, the most predictive of the subject's sequence of responses in the probabilistic task.

Final Q values for fit-to-train simulations yielded a highly significant association between Q value and stimulus/reinforcement condition (Figure 4a of main paper; F[5,340] = 81.0, P $<$ 0.0001)

The rationale for building this model is that the best fitting parameters to participants' training data would be forced to accommodate trial-to-trial adaptions in response to recent reinforcement experiences, such that α_L would reflect a sensitivity to the recency of losses and associated lose-shift performance. We also hypothesized that this parameter would vary by *COMT* genotype.

Fit to Test.

As described in the main paper, we also separately optimized the model's parameters to fit behavioral performance in the *test* (generalization) phase, hypothesizing that a different Q' system is under control over behavior in that phase. The Q' updating equation is identical to that depicted in Eq 5 (Eq 1 of main text). Q' values for each stimulus are computed as a function of reinforcement feedback during the training phase, but with potentially different learning rates than the standard Q system. We then computed the final Q' values associated with each stimulus at the end of training for each set of parameters. The only difference is that the best fitting α values are determined to maximize fit between model and participants' choices made in the test phase, rather than fitting the trial-by-trial behavioral sequences in the training set. Recall that in the fit-to-train case above, Eq. 6 was applied to predict the probability of a participant choosing A over B or C over D or E over F during each trial in the training set. In contrast, here we apply the same equation but to predict the probability that the participant chooses A over C,D,E,F (and all other combinations BC, etc) during the test phase. For example when faced with the novel test pair AC, and the subject chooses A, we compute the probability P_A^{test} as

$$P_A^{test} = \frac{e^{\frac{Q'_{Afinal}}{\beta'}}}{e^{\frac{Q'_{Afinal}}{\beta'}} + e^{\frac{Q'_{Cfinal}}{\beta'}}}, \quad \textbf{[8]}$$

where $Q'(final)$ values reflect the final Q values computed at the end of training, given the current set of α' and β' parameters. Q' values are assumed to not change as a function of test trials (given that no feedback is administered during test). We then found the best fitting parameters $\alpha_{G'}$, α'_L and β' of the Q' system to maximize the likelihood of the generalization test phase choices under the model.

$$LLE(test) = log(\prod_{test} P(test)_{i^*,test}), \quad \textbf{[9]}$$

where $i^*, test$ denotes the subjects choice in each test trial. As for the fit-to-train data, for each subject, the best fit parameters are those associated with the maximum LLE value and are, by definition, the most predictive of the subject's choices in the test phase of the probabilistic task. In sum, this procedure allows us to determine the parameters of a (putative BG) system that learns from reinforcement during the training phase, but only comes to dominate behavioral output in the test phase.

The reasoning for this separate fit to test phase data is that parameter values obtained from fitting the model to the training phase (as above) can capture working memory for the recency of outcomes. In contrast, fitting test phase data can more purely capture a (putative BG) system that had integrated reinforcement values over multiple trials, and is not subject to trial-to-trial recency effects (since there was no feedback in the test phase). We therefore compared learning rates of the fit-to-test procedure to determine whether these would vary by *DARPP-32* and *DRD2* genotype. Larger α' values would indicate that participants are relatively more sensitive to the most recent reinforcement experiences at the end of the training set (just prior to test), whereas smaller values indicate integration of probabilities over multiple training trials.

Final Q values for fit-to-test simulations, yielded a highly significant association between Q value and stimulus/reinforcement condition ($F[5,340] = 48.0$, $P < 0.0001$)

Q Learning Fits.

Table 2 shows mean parameter values across all subjects. Table 3 shows mean LLE values for each subject for data fit to train and test. Note that LLE's are higher when comparing model-to-data fits in the training compared to test phase, simply because there are more training than test trials and therefore greater summed error across trials (similarly, some subjects may have deceptively higher or lower LLE's in fit-to-train as a result of performing more or less training trials before reaching performance criteria.) To provide a more interpretable fit, we calculated pseudo-R^2 values, defined as $(LLE - r)/r$, where r is the log likelihood of the data under a model of purely random choices (p = 0.5 for all choices) (*17,18*). The resulting pseudo-R^2 statistic reveals how well the model fits the data compared to a model predicting chance performance, and is independent of the number of trials to be fit in each set.

To further motivate the need for a separate system that dictates behavior during test, we computed how well the standard fit-to-train Q value learning algorithm can account for choices in the test phase. That is, we applied final Q values from the fit-to-train procedure and computed LLE and pseudo-R^2 on the subsequent test data. We found that indeed, the fit of training data to test phase data was substantially poorer (pseudo-R^2 = 0.16), only half as good as the fit-to-test procedure. That a second (putative BG) system is at play for test choices is highly supported by the reliable associations between these new Q' learning parameters that best fit test responses and the striatal genes.

Additional Q Learning Results.

To provide additional validation of both the Q learning approach and our assumptions about the task, we regressed choose-A and avoid-B performance against final Q' values for each of the stimuli, A,B,C,D,E, and F (all entered in the regression simultaneously, so that any effect of a single Q' value controls for effects of other stimuli). As expected, better choose-A performance was associated with relatively higher Q'_A values (F[1,61] = 7.62, P = 0.0076); this relationship not seen for Q' values of any other stimulus (all P's > 0.4, with the exception of the next-most positive stimulus C, P = 0.065). Similarly, better avoid-B performance was associated with lower Q'_B (F[1,61] 8.36, P = 0.0053), with no effect of Q' values for any other stimulus (all P's > 0.15).

In the main paper we reported that smaller $\alpha_{G'}$ values are associated with better choose-A performance, whereas smaller $\alpha_{L'}$ were associated with better avoid-B performance, supporting the idea that slow integration is necessary for probabilistic generalization. Additional analyses revealed that while there was no overall difference between α_G and $\alpha_{G'}$ F[1,68]= 1.0), this difference became apparent in subjects who successfully generalized positive reinforcement values. That is, $\alpha_{G'}$ was relatively smaller than α_G with increasing choose-A test performance (F[1,67]=3.5, P =.06). There was a similar, albeit nonsignificant, trend for relatively smaller $\alpha_{L'}$ than α_L with increasing avoid-B performance (F[1,67] = 2.2, p=.14). We contend that subjects who did not show lower $alpha'$ than $alpha$ values were overly reliant on recent reinforcement outcomes in the test phase (putatively represented in PFC) and therefore were not successful at generalization.

We also showed that the *DARPP-32* gene modulates slow integration of positive values, supporting discrimination between subtly different reward values (Figure 5*b* of main paper). Here we present converging evidence for this idea. We analyzed the final Q' values for each participant. We then asked whether Q' values in the test phase showed enough fidelity to reliably discriminate between different positive (80, 70 and 60% reward probability) and negative (40, 30 and 20%) values. For positive values, there was a main effect of reinforcement probability (F[2,134] = 17.5, p$<$.0001), such that higher probabilities were associated with significantly higher Q' values. Notably, this effect interacted with *DARPP-32* genotype (F[2,134] = 11.4 P < 0.0001), such that only A/A homozygotes successfully discriminated between Q' values of positive stimuli (SI Fig. 8). This finding confirms that a low $\alpha_{G'}$ in A/A participants allowed these individuals to discriminate between subtly different positive values (consistent with the depiction in Figure 5b of the main paper). No such interaction was observed for either *COMT* or *DRD2* genes (F[2,34] = 0.6 and 0.1, respectively). For negative values, there was again a main effect of condition (F[2,134] = 6.9,

p=.001), but this did not interact with *DARPP-32* (F[2,134] = 1.25, ns), *DRD2* (F[2,134]=0.4) or *COMT* (F[2,134]=.4) genotypes.

Extended BG-OFC Q Learning Model

The purpose of the separate fit-to-train and fit-to-test models was to show that two systems (putatively BG and OFC) learn in parallel during the training phase of the task. Fitting to training data can best capture a system adapting on a trial-to-trial basis, whereas fitting to test best captures the accrued reinforcement values over all of training. This led us to explore an extended model that has two separate systems with different learning rates in the training phase, including a working memory system that decays with time, and a BG system. The two systems contribute to a single Q value for each stimulus, but which is updated by two different learning rates. If our assumptions are correct, then the best fitting learning rate to the decaying (PFC) system should be substantially higher than that of the non-decaying (BG) system. Because the higher learning rate would dominate Q value updates, this would lead to the PFC dominating Q updates early, while the slower BG would dominate updates later (once the PFC system has decayed).

In the combined model, Q values are computed as follows. Rather than using separate learning rates for losses and gains, we instead use separate learning rates for two systems, maintaining the same total number of free parameters:

$$Q_i(t+1) = Q_i(t) + \alpha_{BG}[r(t) - Q_i(t)] + \alpha_{OFC}[r(t) - Q_i(t)]e^{-0.5t}. \qquad \textbf{[10]}$$

The term $e^{-0.5t}$ is introduced to simulate a decay of working memory strategies as training progresses (eg, SI Fig. 6). We used a constant decay for all subjects to eliminate the need to search simultaneously across multiple parameters.[1] The BG component of the Q value does not decay with time, because BG learning becomes more habitual with increased training (*19*). Thus although our fit-to-training and fit-to-test procedures assume two systems where only PFC governs behavior during training and only BG governs behavior at test [similar to the binary use of two systems in (*20*)], here we impose a more soft constraint whereby both systems contribute to behavior, but with relatively greater use of BG and less of PFC as trials progress.

Consistent with our overall hypothesis, we found that the resulting best fit α_{OFC} was on average

[1]The 0.5 factor was chosen to approximate the time course of decay in lose-shift performance with trials. However, other simulations revealed that the exact value of this parameter does not change the patterns or significance of the findings reported here.

twice as high as that of α_{BG} (Table 4; F[1,68] = 10.71, p= 0.0017). The combined BG/OFC model also yielded a decent fit to the training phase data, although its fits to behavior, especially to test phase data, were not as good as the Gain-Loss Q learning model (Table 5). Moreover, unlike the Gain-Loss model, this model is not able to account for differential effects of losses versus gains or associated genetic effects, as it does not incorporate different learning rates for gains and losses. To do so would require incorporating separate gain and loss terms for each of the BG and OFC systems, which would amount to 5 free parameters (including β), which is less parsimonious and computationally intractable given the large search space and exponential effects of combining parameters. Although a nonlinear search optimization algorithm is possible in principle, these are subject to local maxima and possible interactions between parameters, making it much more difficult to find clear genetic/parameter dissociations.

We further analyzed whether individual subjects' data were better fit by the Gain-Loss or BG/OFC Q learning models. These models both include the same number of parameters, but the two learning rate parameters are allowed to vary either for gains versus losses or for fast (decaying) vs slow learning systems. We hypothesized that when fitting to train data, the BG/OFC model would provide a better fit than the Gain-Loss model with increasing *COMT* met expression, since these individuals behaviorally showed increased lose-shift (working memory) effects that decayed with time on task, which would be captured with the BG/OFC model. We computed the per-subject relative difference in pseudo-R^2 between BG/OFC and Gain-Loss Q models for the fit-to-train simulations. This difference measures the degree to which having an explicit mechanism for increased working memory early in training can improve fit to behavior relative to a model that has only different learning rates for gains and losses. We found that this difference indeed correlated with increasing met allele expression (r(68) = 0.26, one-tailed P = .015). This result supports the notion that increasing met allele expression (and associated PFC DA) requires a parameter to capture explicit working memory contributions, and relatively less differentiation between losses and gains.

Similarly, we hypothesized that fit-to-test procedure would yield a better fit for the Gain-Loss model than the BG-OFC model, dependent on the *DARPP-32* gene. Recall that the A/A genotype was associated with relatively better Go than NoGo test generalization. We therefore reasoned that these participants' test performance would be relatively better fit by a model that allowed separate learning rates for loss and gain (Gain-Loss). Indeed, the pseudo-R^2 fit for the Gain-Loss model was relatively higher than that of the BG-OFC model for A/A participants compared to G carriers

(t(68) = 1.7, one-tailed P = 0.04). There was no such effect for the *DRD2* gene ($p > 0.4$).

References

1. Asaad WF, Rainer G, Miller EK (1998), *Neuron* 21:1399-1407.
2. Raichle ME, Fiez JA, Videen TO, MacLeod AM, Pardo JV, Fox PT, Petersen SE (1994), *Cerebral Cortex* 4:8-26.
3. Laming D (1979), *Acta Psychologica* 43:199-224.
4. Judd CM, McClelland GH (1989), *Data Analysis, A Model-Comparison Approach* (Harcourt Brace Jovanovich, Orlando, FL).
5. Tunbridge EM, Harrison PJ, Weinberger DR (2006), *Biol Psychiatry* 60:141-151.
6. Roberts AC, De Salvia MA, Wilkinson LS, Collins P, Muir JL, Everitt BJ, Robbins TW (1994), *J Neurosci* 14:2531-2544.
7. Saunders RC, Kolachana BS, Weinberger DR (1998), *Nature* 393:169-171.
8. Schmitz Y, Benoit-Marand M, Gonon F, Sulzer D (2003), *J Neurochem* 87:273-289.
9. Sesack SR, Hawrylak VA, Matus C, Guido MA, Levey AI (1998), *J Neurosci* 18:2697-2708.
10. Di Chiara G, Loddo P, Tanda G (1999), *Biol Psychiatry* 46:1624-1633. .
11. Jackson ME, Moghaddam B (2004), *J Neurochem* 88:1327-1334.
12. Frank MJ, Claus ED (2006), *Psychol Rev* **113**:300-326.
13. Watkins CJCH, Dayan P (1992), *Machine Learning* 8:279-292.
14. Sutton RS, Barto AG (1998), *Reinforcement Learning: An Introduction.* (MIT Press, Cambridge, MA).
15. Frank MJ, Seeberger LC, O'Reilly RC (2004), *Science* 306:1940-1943. .
16. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006), *Nature* 442:1042-1045.
17. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006), *Nature* 441:876-879.

18. Camerer C, Ho T-H (1999), *Econometrica* 67:827-874.

19. Yin HH, Knowlton BJ (2006), *Nat Rev Neurosci* 7: 464-476.

20. Daw ND, Niv Y, Dayan P (2005), *Nat Neurosci* 8:1704-1711.

Table 1. Reaction times for *COMT* and *DRD2* genotypes.

Genotype	Choose-A	Avoid-B	Post-correct	Post-error
DRD2				
C/C	969 (100)	1,279 (134)	801 (45)	882 (63)
C/T	1,105 (106)	1,186 (94)	1,008 (80)	1,168 (103)
T/T	1,046 (118)	1,076 (94)	1,120 (156)	1,211 (185)
COMT				
val/val	1,224 (143)	1,337 (122)	1,040 (136)	1,135 (157)
val/met	1,042 (107)	1,084 (84)	1,032 (80)	1,165 (112)
met/met	993 (104)	1,225 (129)	945 (127)	1,093 (144)

No significant differences were observed for raw RT's, but only in the relative measures described in the main text (For *DRD2*: Avoid-B compared with Choose-A RTs; and for *COMT*: Post-error compared with post-correct RTs). Values in parentheses reflect standard error.

Table 2. Mean best fitting parameter values for Gain-Loss Q model.

α_G	α_L	β	$\alpha_{G'}$	$\alpha_{L'}$	β'
0.36 (0.24)	0.14 (0.21)	0.29 (0.18)	0.31 (0.36)	0.22 (0.33)	0.2 (0.24)

Standard deviations are in parentheses.

Table 3. Per-subject mean LLE and pseudo-R^2 values for model fit to data.

	Train	Test
LLE	-72.75 (63.64)	-33.87 (32.0)
pseudo-R^2	0.327 (0.19)	0.324 (0.24)

Standard deviations are in parentheses. Train: parameters optimized to fit training phase trial-by-trial data as a function of feedback and time. Test: parameters optimized during the learning phase in order to fit subsequent test phase performance.

Table 4. Mean best fitting parameter values to BG-OFC Q learning model.

α_{BG}	α_{OFC}	β
0.17 (0.15)	0.33 (0.39)	0.32 (0.23)

Standard deviations are in parentheses.

Table 5. Per-subject mean LLE and pseudo-R^2 values for BG-OFC Q learning model.

	Train	Test
LLE	-74.4 (62.5)	-35.13 (32.2)
pseudo-R^2	0.309 (0.18)	0.29 (0.25)

Standard deviations are in parentheses.

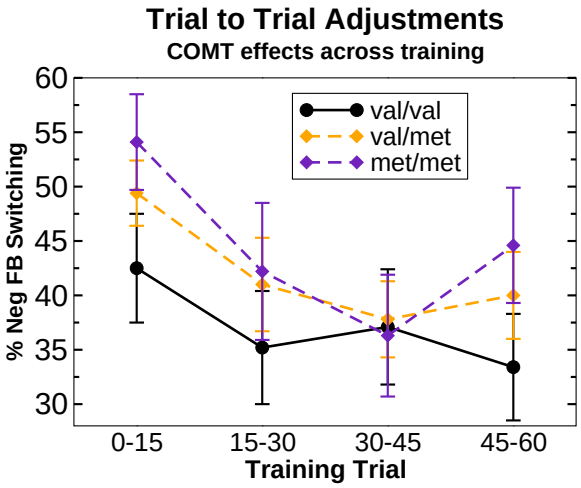

Figure 6: ***COMT* effects on trial-to-trial adjustments as a function of training trial.** Effects of negative feedback on subsequent switching were strongest in early training trials, and decreased as training progressed. Error bars reflect standard error.

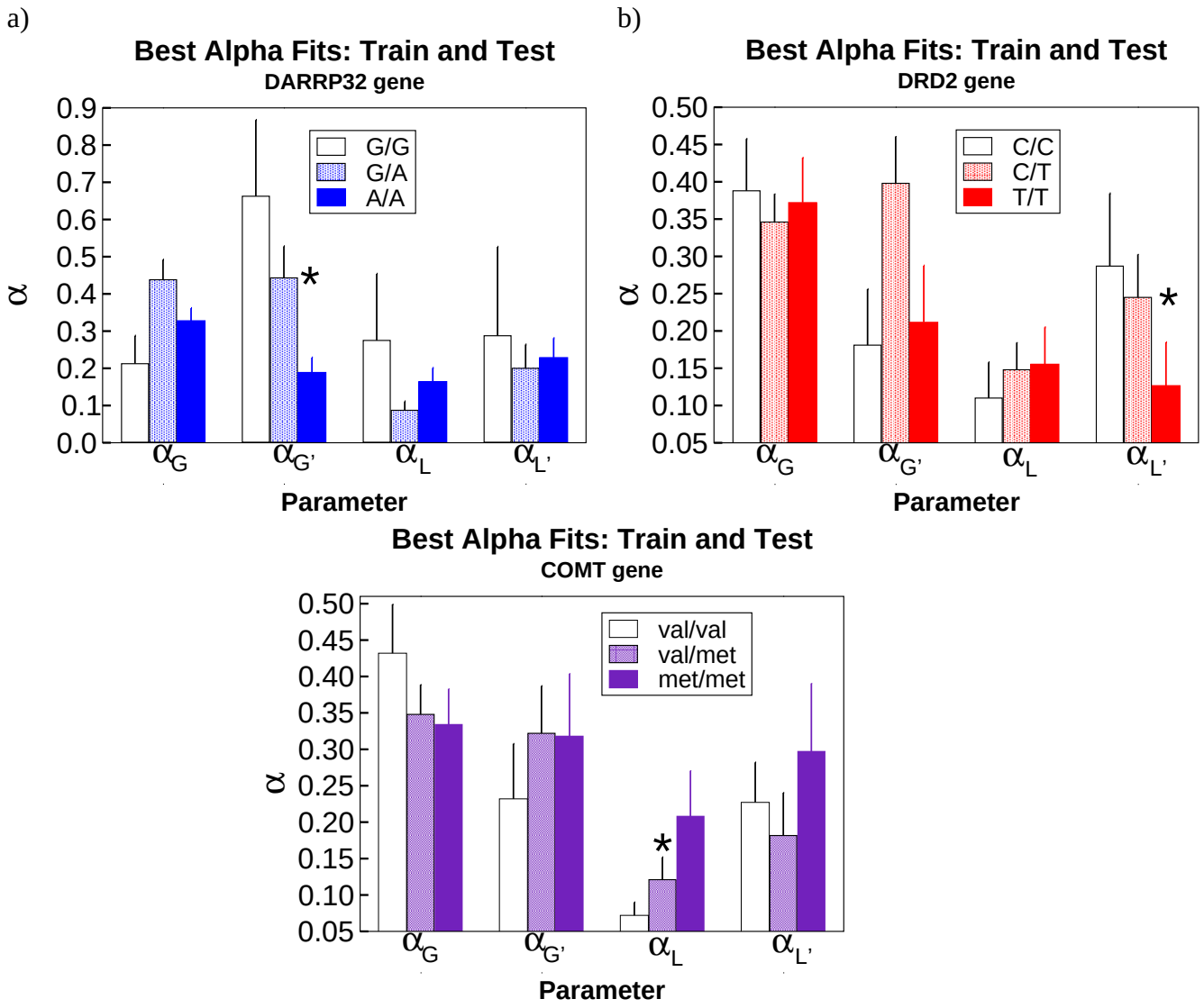

Figure 7: Gene dose effects on Q learning parameters. **a)** DARPP32 gene. A/A homozygotes had smaller α'_G values (accounting for generalization of probabilistically integrated positive outcomes). **b)** *DRD2* gene. T/T homozygotes had smaller α'_L values (accounting for generalization of probabilistically integrated negative outcomes). **c)** *COMT* gene. Val/Val homozygotes had smaller α_L (accounting for reduced trial-to-trial modification of reinforcement values). There were no gene effects on α_G or α'_G.

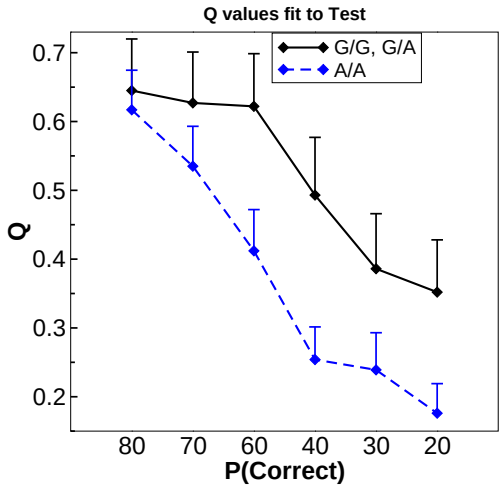

Figure 8: *DARPP-32* effects on final Q' values from fit-to-test simulations. A/A homozygotes could differentiate between positive stimulus values (80, 70 and 60%), whereas G carriers showed similar positive Q' values for these stimuli. No such *DARPP-32* effect was observed for discriminating between negative values. Error bars reflect standard error.