The neurogenetics of exploration and exploitation Supplementary Material

Michael J. Frank[1*], Bradley B. Doll[1], Jen Oas-Terpstra[2] & Francisco Moreno[2]

[*]Corresponding author: michael_frank@brown.edu

[1] Depts of Cognitive & Linguistic Sciences and Psychology,
Brown Institute for Brain Science, Brown University

[2] Dept of Psychiatry, University of Arizona

Additional Task Results: Probability-magnitude bias

As noted in the main text, the CEVR condition was included to compare with CEV as a measure of probability-magnitude bias (PM-bias = CEVR - CEV). We found PM-bias to be positive across all groups ($p < .001$; Figure S1), as participants avoided responding early in CEVR due to the low reward probability at that time, consistent with loss-aversion (*1*). Supporting this interpretation, DRD2 T/T carriers, who showed enhanced NoGo learning as assessed by IEV_{diff} above, also showed relatively greater PM-bias than C carriers (F(1,66) = 4.7, p =.03; Figure S1b). Their RTs in the last block of CEVR were also significantly slower (F[1,66] = 5.7, p = .02). Both IEV_{diff} and PM-bias were also elevated in non-medicated Parkinson's patients (*2*), consistent with their performance in other learning paradigms (*3*–*5*).

Although on average participants showed positive PM-bias, we reasoned that those with enhanced sensitivity to reward magnitudes would exhibit less of a bias. Based on neurocomputational models and physiological data (*6*–*9*), we posited that magnitude representations are maintained in orbitofrontal cortex (OFC), a brain area that is particularly sensitive to COMT effects (*10*). We therefore predicted that met allele carriers would properly incorporate reward magnitudes into their expected value computations and would therefore show less of a probability bias. There was only weak evidence for such a finding in the last quarter of trials (F(1,67) = 2.5, p=.12; Figure S1); this effect was significant however when measured across all trials (F(1,67) = 5.2, p =.026).

Note that the above genetic interpretations rely on two components to PM-bias: a putative striatal-NoGo bias that learns from high frequency of non-rewards and a putative prefrontal representation of reward magnitudes that can counteract this bias. This assumption implies that DRD2 and COMT effects independently contribute to PM-bias. Supporting this conclusion, when both D2 and COMT genotypes were entered into the statistical model, D2 effects on PM-bias remained significant (F(1,65) = 5.5, p =.02), and the COMT effects approached significance (F(1,65) = 3.3, p=.07). (Across all trials, both D2 and COMT effects were significant each controlling for the other; p $<$ .02.) Indeed, D2 and COMT effects are additive: individuals with superior striatal D2 genetic function but poor prefrontal genetic function (T-val) show by far the strongest PM-bias, whereas C-met individuals with worse striatal D2 but better prefrontal function show effectively no PM-bias (Figure S1d). This result suggests that PM-bias emerges both from a striatal-dependent NoGo learning (loss avoidance), and prefrontal-dependent sensitivity to high magnitude rewards.

Alternative Kalman Filter Exploration Model

As mentioned above, the primary model assumed that subjects track the probability of obtaining a reward prediction error separately for fast and slow responses. However, it is also possible that they would track the magnitude of such prediction errors (or the magnitudes of raw rewards), and the uncertainty thereof. In this case, the beta distribution would be inappropriate, and instead Gaussian distributions $N(\mu, \sigma^2)$ can be used to represent the mean expected values for each response. The Gaussian has the probability density function,

$$\frac{1}{\sigma\sqrt{(2\pi)}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Kalman filter (*11*) is a Bayesian algorithm that can track the mean values of arbitrary quantities (see Welch & Bishop 1995 (*12*) for an introduction and derivation). The filter is in many ways similar to the classical Rescorla-Wagner delta rule (*13*) in that values are learned as a function of the difference between prior estimates and observed outcomes, but instead of representing only the best guess for the expected value of a quantity, it represents an entire distribution of guesses, including the mean "best" guess, and the uncertainty about it (*14, 15*). We represented the mean reward values for fast and slow responses, and updated these values as follows for the selected action (fast or slow):

$$\mu_{s,a}(t+1) = \mu_{s,a}(t) + k_{s,a}(t)[Rew_{s,a}(t) - \mu_{s,a}(t)]$$

(In an alternative model we represented the mean reward prediction error values, replacing $Rew_{s,a}(t)$ with $\delta_{s,a}(t)$ above).

where k is the Kalman gain (effective learning rate),

$$k_{s,a}(t) = \frac{\sigma_{s,a}(t)^2}{\sigma_{s,a}(t)^2 + \sigma_{rew}^2},$$

$\sigma_{s,a}(t)$ is the standard deviation of the Gaussian distribution, and σ_{rew} is the noise in the reward signal. The posterior variance of the action taken is

$$\sigma_{s,a}(t+1) = [1 - k_{s,a}(t)]\sigma_{s,a}(t),$$

such that the uncertainty reduces with experience. This mechanism allows the model to estimate the mean magnitude of rewards, or of reward prediction errors, for fast and slow responses as a function of experience.

Note the Kalman filter requires additional parameters to initialize the means and variances of the Gaussians (in contrast to the beta distributions which are initialized to be uniform). To reduce the number of free parameters and to compare with the beta implementation, we initialized $\mu_{s,a}(1) = 0$ (i.e., the initial expected magnitude of prediction error is 0), and we initialized $\sigma_{s,a}(1) = \sigma_{rew}$ to be the standard deviation of the actual reward vector, providing a non-arbitrary objective initial uncertainty without requiring an additional parameter (other reasonable values produce similar results). This implies that the Kalman gain (learning rate) begins with a value of 0.5, and then decreases according to the Bayesian updating rule as a function of uncertainty. As the standard deviation decreases, so does the gain. Once the means and standard deviations are derived, we then applied them in an identical fashion to that described above for beta implementation (using ρ to scale the differences in the means to adapt RT in the exploitation model, and ϵ to scale the differences in standard deviations to drive exploration).

While this model did not fit the behavioral data as well as the beta distribution model, all genetic effects showed identical patterns to that described in the main paper. For the exploit part of the model, when estimating reward prediction error magnitudes, the DARPP-32 effect on relative α_G to α_N was significant at p = .02; the DRD2 was significant at p = .016, and the COMT gene dose effect on uncertainty-based exploration was significant at p = .007. Similarly, when estimating raw reward magnitudes, all effects held (DARPP32, p = .03; DRD2, p = .038, COMT gene-dose, p = .01). No other parameters were modulated by genotype (p’s $>$ 0.2).

Explore Model: Fits and Comparison to Foil Models

Relative to the base exploitation model, the uncertainty-based explore model provided a better fit (according to AIC) overall. Notably, the improvement in fit was significant in met allele carriers

(F[1,49] = 5.1, p = .029), but actually provided a poorer fit in val/val participants (F[1,18] = 7.5, p = .01). This was reflected by a significant interaction, such that the change in model fit due to the addition of an uncertainty-exploration term depended on COMT genotype (F[1,67] = 6.0, p = .016; Figure S3).

This was not the case for either the Sutton (1990) (*16*) exploration bonus model or the lose-switch model, neither of which showed a fit improvement relative to the base exploitation models once penalized for additional parameters, and both clearly inferior to the uncertainty model. (There were also no effects of COMT on either of these models' parameters).

The regression to the mean model was motivated by behavioral results showing large regression to the mean effects in this task (Figure S2), as described previously (*2*). This model did not clearly fit better than the base model however. The reverse-momentum model, which provided a more local oscillation rule that allowed RT momentum to build in one direction before reversal, did improve fit. However, when comparing this model fit to that of the uncertainty-explore model, there was again a significant interaction by COMT genotype (F[1,67] = 4.7, p = .03). This was due to the uncertainty-explore model providing a better fit than the reverse-momentum model only in met/met participants (F[1,49] = 5.4, p = .02). Moreover, the COMT gene-dose effect held when comparing ϵ relative to either ξ (F[1,67] = 3.8, p = .05), κ (F[1,67] = 7.7, p = .007), or γ (F[1,67] = 6.4, p = .01; Figure 8b,c of main text). (In this analysis, z-scores were computed for each parameter so that they can be compared in the same metric). The same gene-dose relationships help when both uncertainty explore and one of the alternative models of trial-to-trial dynamics were included in the same model.

Kalman Filter Model Results

While this model did not fit the behavioral data as well as the beta distribution model, all genetic effects showed identical patterns to that described in the main paper. For the exploit part of the model, when estimating reward prediction error magnitudes, the DARPP-32 effect on relative α_G to α_N was significant at p = .02; the DRD2 was significant at p = .016, and the COMT gene dose effect on uncertainty-based exploration was significant at p = .007. Similarly, when estimating raw reward magnitudes, all effects held (DARPP32, p = .03; DRD2, p = .038, COMT gene-dose, p = .01). No other parameters were modulated by genotype (p's $>$ 0.2).

Generative Model

While the presented models provide reasonable fits to data, it is also important to show that a reinforcement model is adaptive when it runs on its own. To this end, we fixed the model parameters to reasonable values (K = 1500, λ = 0.2, ϵ = 3000, α_G = α_N = 0.3, ν = 0.2; ρ = 1000; Noise = 2000). The model selected its own responses with these parameters, and was rewarded with the same reward probability and magnitude functions used with participants. 70 runs of the model were simulated, and it clearly produced the adaptive pattern of results as shown in Figure S7.

References

1. A. Tversky, D. Kahneman, *Science* **211**, 453 (1981).

2. A. A. Moustafa, M. X. Cohen, S. J. Sherman, M. J. Frank, *Journal of Neuroscience* **28**, 12294 (2008).

3. M. J. Frank, L. C. Seeberger, R. C. O'Reilly, *Science* **306**, 1940 (2004).

4. M. J. Frank, J. Samanta, A. A. Moustafa, S. J. Sherman, *Science (New York, N* **318**, 1309 (2007).

5. R. Cools, L. Altamirano, M. D'Esposito, *Neuropsychologia* **44**, 1663 (2006).

6. M. J. Frank, E. D. Claus, *Psychological review* **113**, 300 (2006).

7. M. R. Roesch, C. R. Olson, *Science (New York, N* **304**, 307 (2004).

8. N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, R. J. Dolan, *Nature* **441**, 876 (2006).

9. J. P. O'Doherty, *Annals of the New York Academy of Sciences* **1121**, 254 (2007).

10. M. Slifstein, *et al.*, *Molecular psychiatry* **13**, 821 (2008).

11. R. E. Kalman, *Transactions of the ASME – Journal of Basic Engineering* pp. 35–45 (1960).

12. G. Welch, G. Bishop, An introduction to the kalman filter, *technical report* ,, University of North Carolina at Chapel Hill Dept of Computer Science (1995).

13. R. A. Rescorla, A. R. Wagner, *Classical Conditioning II: Theory and Research*, A. H. Black, W. F. Prokasy, eds. (Appleton-Century-Crofts, New York, 1972), pp. 64–99.

14. P. Dayan, S. Kakade, P. R. Montague, *Nat Neurosci* **3 Suppl**, 1218 (2000).

15. J. K. Kruschke, *Learning & behavior* **36** (2008).

16. R. S. Sutton, *Proceedings of the Seventh International Conference on Machine Learning*, B. W. Porter, R. J. Mooney, eds. (Morgan Kaufmann, Palo Alto, CA, 1990), pp. 216–224.

Cond/Group	n	DEV	CEV	IEV	CEVR
COMT (met)	50	1810 (72)	2024 (71)	2353 (83)	2212 (91)
COMT (val/val)	19	1703 (142)	1870 (139)	2465 (170)	2467 (110)
DRD2 (T/T)	38	1798 (88)	1949 (90)	2437 (96)	2404 (104)
DRD2 (C)	31	1759 (96)	2018 (93)	2319 (122)	2132 (98)
DARPP-32 (T/T)	37	1697 (94)	2013 (83)	2414 (100)	2207 (107)
DARPP-32 (C)	27	1927 (92)	1857 (107)	2377 (133)	2365 (112)

Table 1: Response times (ms) in each task condition across all trials, broken down into genotypes for each polymorphism. Values reflect mean (standard error).

Param/Model	Base	Uncert	Regress	L-Switch	Rev-Mom	Exp-Bonus
K	1416.0 (48.0)	1357.5 (47.2)	1341.9 (49.8)	1407.6 (48.4)	1378.0 (47.5)	1406.3 (48.2)
λ	0.325 (.016)	0.34 (.016)	0.36 (.017)	0.33 (.016)	0.35 (.016)	0.33 (.016)
α_G	0.25 (.04)	0.24 (.045)	0.247 (.04)	0.25 (.04)	0.25 (.04)	0.25 (.04)
α_N	0.27 (.05)	0.28 (.05)	0.266 (.05)	0.27 (.05)	0.27 (.05)	0.27 (.05)
ρ	444.8 (51.7)	436.0 (53.5)	437.7 (50.6)	434.7 (51.8)	429.3 (50.6)	443.2 (51.4)
ν	0.11 (.01)	0.11 (.01)	0.11 (.01)	0.11 (.01)	0.11 (.01)	0.11 (.01)
ϵ	–	2233.9 (290.9)	–	–	–	–
ξ	–	–	51.0 (7.8)	–	–	–
κ	–	–	–	10.5 (2.9)	–	–
γ	–	–	–	–	40.5 (4.9)	–
θ	–	–	–	–	1.8 (0.2)	–
ζ	–	–	–	–	–	4.66 (2.0)

Table 2: Mean best-fitting parameters for different models. Base: Exploitation model (ie. no parameters for trial to trial adaptation). Uncert: uncertainty-based exploration, with parameter ϵ. Regress: regression to the mean, with parameter ξ; L-Switch: Lose-switch, with parameter κ. Values reflect mean (SE).

Genotype	Model	n params	SSE (x 1e3)	AIC
val/val				
	Just K	1	15696 (20328)	3228.9 (26.7)
	K, λ	2	68735 (3874)	3091.6 (12.8)
	Exploit1	4	64460 (4239)	3080.9 (14.1)
	Exploit2	5	58307 (4016)	3062.8 (13.6)
	Exploit3	6	54371 (3838)	3050.7 (13.8)
	Uncert	7	54112 (3827)	3051.8 (13.8)
	L-Switch	7	54184 (3779)	3052.2 (13.7)
	Exp-Bonus	7	54321 (3851)	3052.5 (13.9)
	Regress	7	540738 (3857)	3051.5 (14.0)
	Rev-Mom	8	53074 (3731)	3050.1 (13.7)
	Kalman	7	54107 (3777)	3052.1 (13.6)
met				
	Just K	1	12578 (8929)	3138.9 (39.5)
	K, λ	2	66438 (2636)	3030.7 (36.2)
	Exploit1	4	64746 (2574)	3029.6 (36.0)
	Exploit2	5	58157 (2276)	3010.7 (35.4)
	Exploit3	6	54706 (2073)	3001.3 (35.1)
	Uncert	7	53773 (2065)	2999.7 (35.0)
	L-Switch	7	54650 (2071)	3003.1 (35.1)
	Exp-Bonus	7	54653 (2075)	3003.1 (35.1)
	Regress	7	54253 (2069)	3001.8 (35.2)
	Rev-Mom	8	53775 (2036)	3002.0 (35.0)
	Kalman	7	54927 (2054)	3004.3 (35.2)

Table 3: Model fits for val/val and met carriers. Exploit1 = Reinforcement learning model with α_G and α_N. Exploit2 = Exploit1 + ν. Exploit3 = Exploit 2 + Bayesian (i.e. $\rho \neq 0$). Uncert: uncertainty-based explore ($\epsilon \neq 0$). L-Switch: lose-switch ($\kappa \neq 0$). Exp-Bonus: Sutton (1990) Exploration bonus ($\zeta \neq 0$). Regress: regression to mean ($\xi \neq 0$). Rev-Mom: reverse momentum (γ,$\theta \neq 0$). Kalman: Kalman filter with Normal distributions and uncertainty-based explore. SSE = sum of squared error; AIC = Aikake's Information Criterion. For both SSE and AIC, lower values indicate better fit. n params = number of parameters. Values reflect mean (standard error).

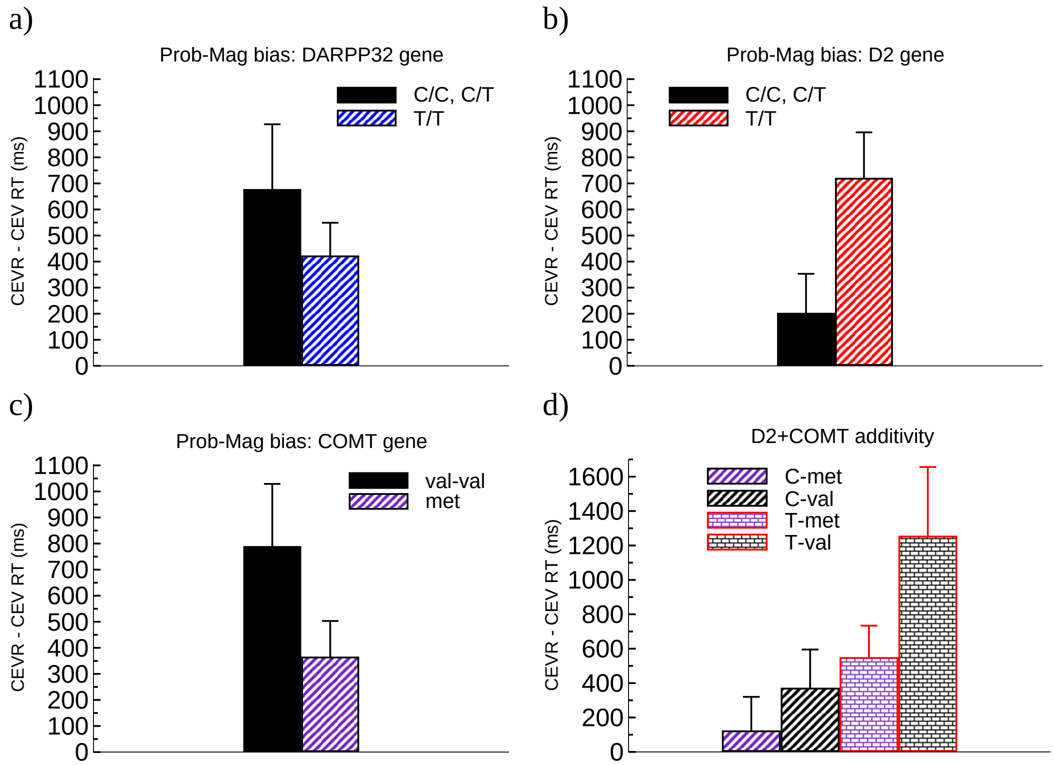

Figure 1: Relative within-subjects biases to prefer high probability over high magnitude, controlling for equal expected value (CEVR - CEV). Values represent mean (standard error) in the last quarter of trials in each condition. **(a-c)** DRD2 and COMT, but not DARPP-32, affected PM-bias. **d)** DRD2 and COMT contributed additively to PM-bias, such that C-met participants with genotype exhibited smallest PM-bias whereas T-val participants exhibited highest PM-bias.

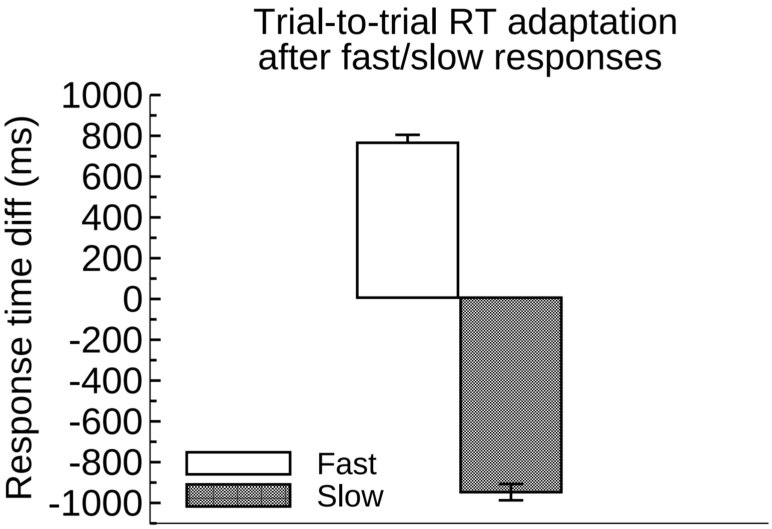

Figure 2: RT changes from one trial to the next reveal regression to the mean effects, whereby prior fast and slow responses are associated with subsequent slowing and speeding, respectively.

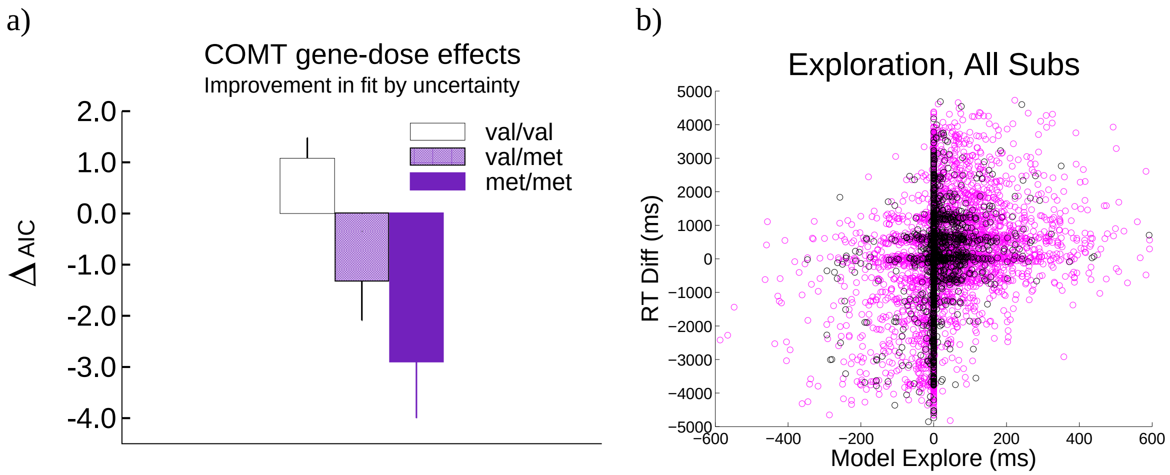

Figure 3: **a)** Improvement in fit (Δ AIC; negative values indicate better fit) afforded by inclusion of the uncertainty-explore term in the model relative to model without exploration. **b)** Scatter plot of all RT swings (change in RT from one trial to the next) against model uncertainty-based exploratory predictions. Met allele carriers are shown in magenta; val/val in black.

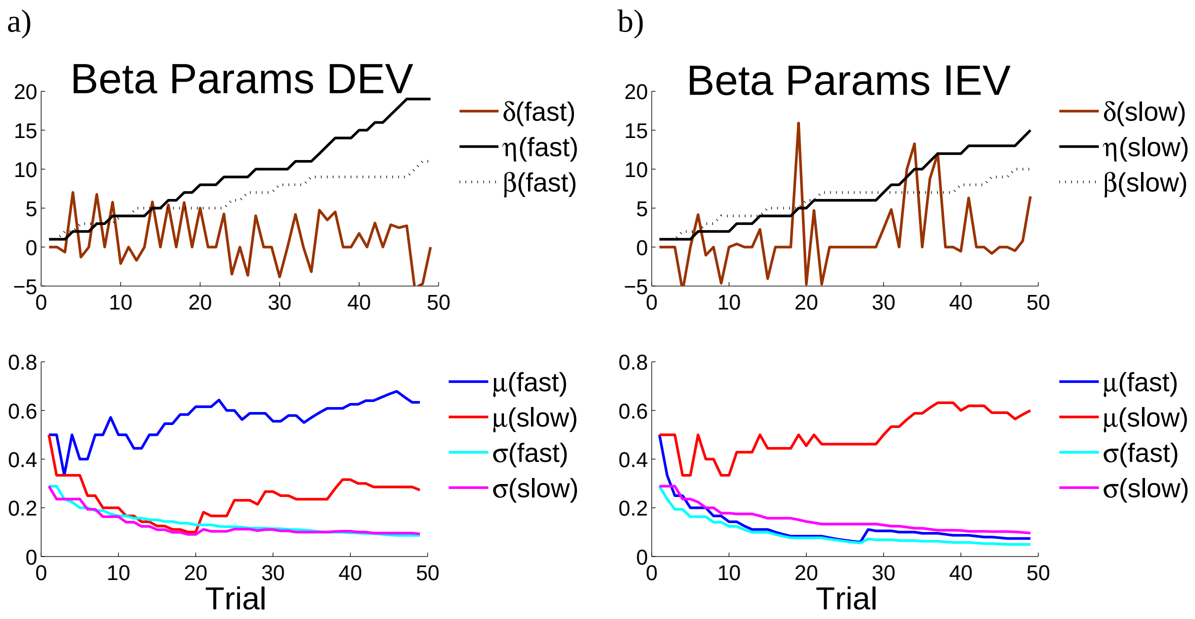

Figure 4: **a), b)** Trajectory of prediction errors (δ) and Beta hyperparameters η and β for a single subject in DEV and IEV. η and β accumulate with evidence obtained on each trial (positive and negative prediction error, respectively), and are used to compute means and variances. δ is scaled to fit in the same axis.

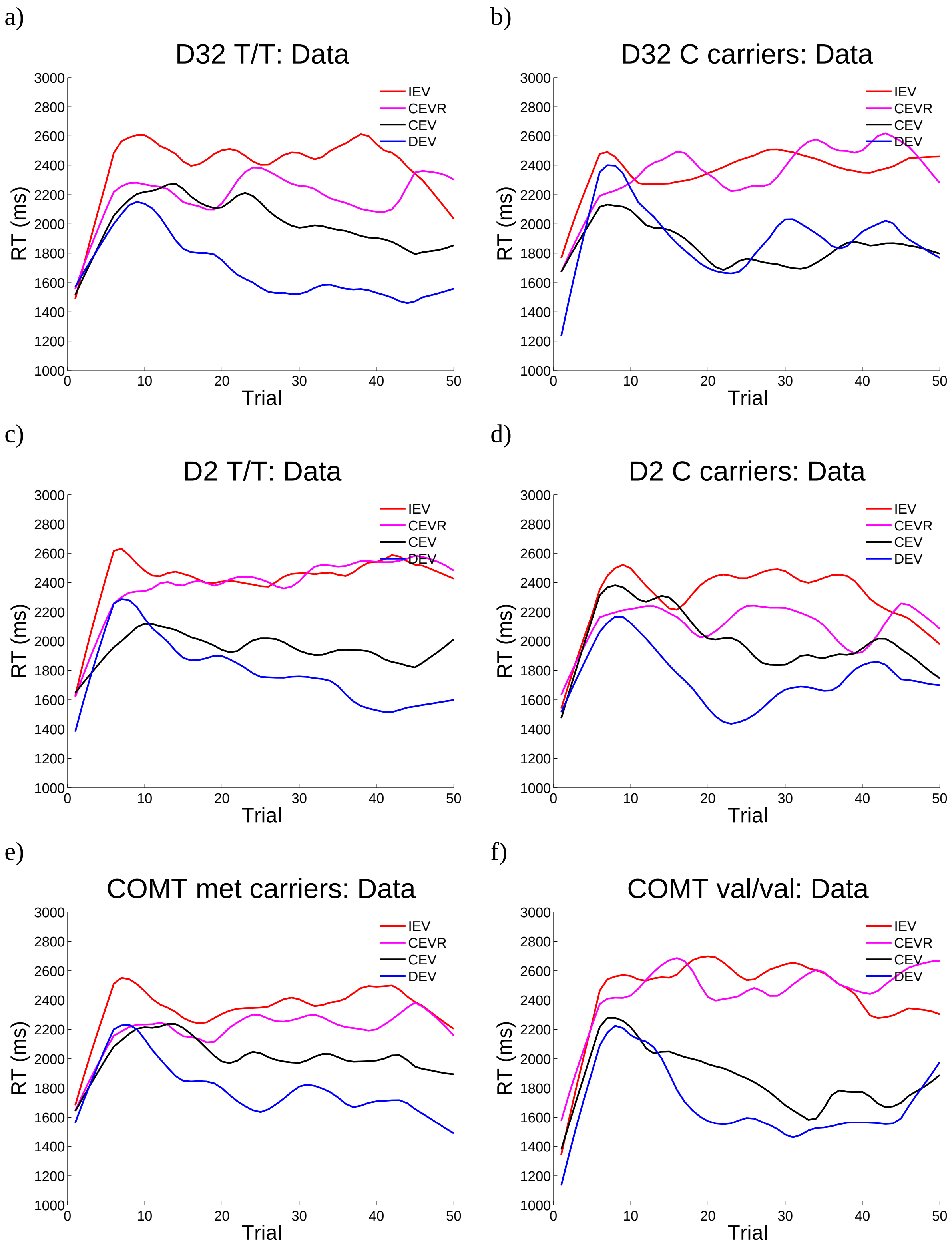

Figure 5: Response times as a function of trial number, same conventions as reported across all subjects in the main text, separated according to genotype.

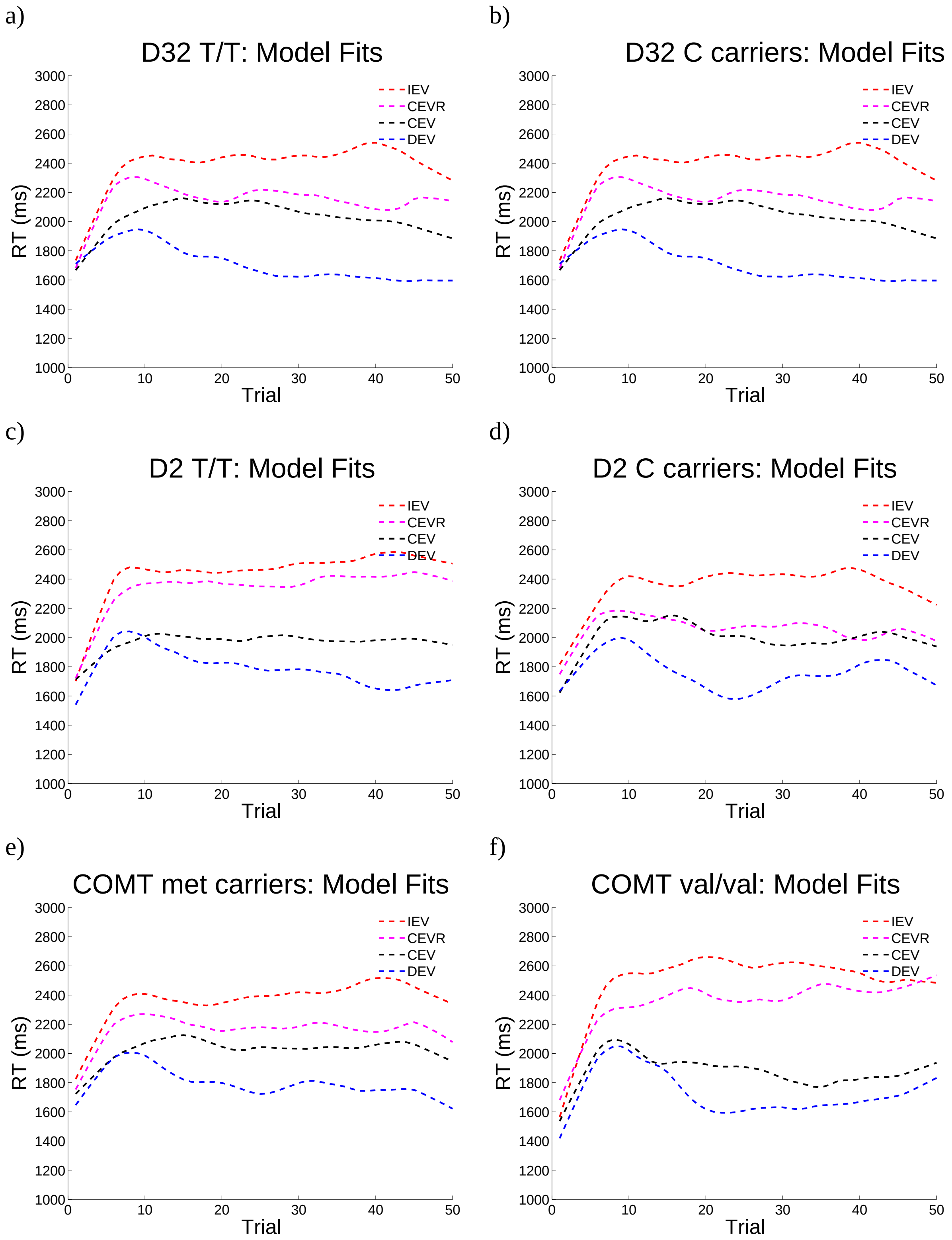

Figure 6: Model fits for each genotype.

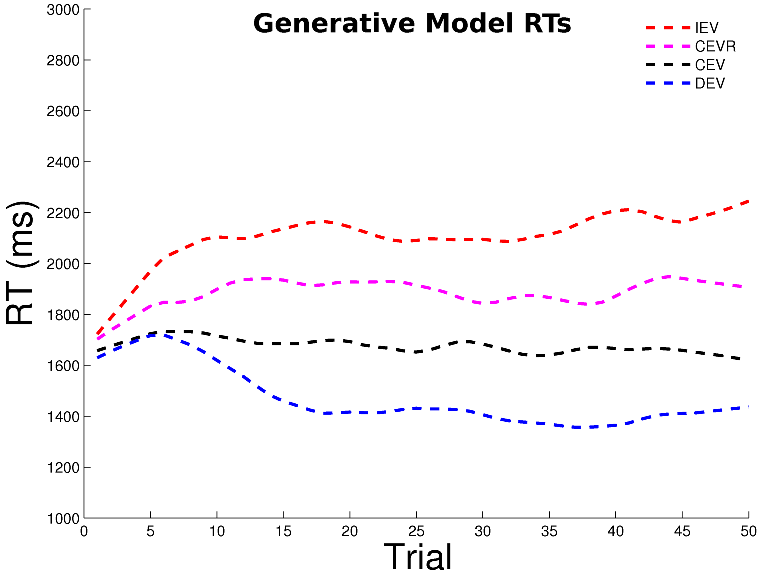

Figure 7: RTs produced by the generative model with fixed parameters across 70 runs, same convention as in the plots of model and subject RTs in the main text.