A Frontal Dopamine System for Reflective Exploratory Behavior

Abbreviated Title: A Frontal Exploratory System

Nathaniel J. Blanco[1], Bradley C. Love[2], Jessica A. Cooper[1], John E. McGeary[3,4], Valerie S. Knopik[3,4], W. Todd Maddox[1]

[1]Department of Psychology, University of Texas at Austin, 78712, United States
[2]Experimental Psychology, University College London, WC1H OAP, UK
[3]Department of Psychiatry & Human Behavior, Warren Alpert Medical School, Brown University, 02903, United States
[4]Division of Behavioral Genetics, Rhode Island Hospital, 02903, United States

Corresponding Author:
Nathaniel J. Blanco
University of Texas at Austin
Department of Psychology
108 E. Dean Keeton Stop A8000
Austin, TX 78712-1043
nathanblanco@gmail.com

Acknowledgments:

This study was funded by the National Institute on Drug Abuse, grant DA032457 to WTM, the Leverhulme Trust grant RPG-2014-075 to BCL, and U.S. Department of Veteran Affairs (VA) shared equipment program to JEM. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs. The authors declare no competing financial interests. We would like to thank Alex Kline, Seth Koslov, and the rest of the Maddox Lab RAs for their help with data collection. We also thank anonymous reviewers for their helpful suggestions.

Abstract

The *COMT* gene modulates dopamine levels in prefrontal cortex with Met allele carriers having lower COMT enzyme activity and, therefore, higher dopamine levels compared to Val/Val homozygotes. Concordantly, Val/Val homozygotes tend to perform worse and display increased (interpreted as inefficient) frontal activation in certain cognitive tasks. In a sample of 209 participants, we test the hypothesis that Met carriers will be advantaged in a decision-making task that demands sequencing exploratory and exploitive choices to minimize uncertainty about the reward structure in the environment. Previous work suggests that optimal performance depends on limited cognitive resources supported by prefrontal systems. If so, Met carriers should outperform Val/Val homozygotes, particularly under dual-task conditions that tax limited cognitive resources. In accord with these a priori predictions, Met carriers were more resilient in the face of cognitive load, continuing to explore in a sophisticated manner. We fit computational models that embody sophisticated reflective and simple reflexive strategies to further evaluate participants' exploration behavior. The Ideal Actor model reflectively updates beliefs and plans ahead, taking into account the information gained by each choice and making choices that maximize long-term payoffs. In contrast, the Naïve Reinforcement Learning (RL) model instantiates the reflexive account of choice, in which the values of actions are based only on the rewards experienced so far. Its beliefs are updated reflexively in response to observed changes in rewards. Converging with standard analyses, Met carriers were best characterized by the Ideal Actor model, whereas Val/Val homozygotes were best characterized by the Naive RL model, particularly under dual-task conditions.

1. Introduction

Effective decision-making requires a balance of exploratory and exploitative behavior (Daw et al., 2006, Cohen et al., 2007, Hills et al., 2015). For example, consider the problem of choosing the best route to work. Routes change over time because of construction, changes in traffic patterns, etc. such that one cannot be certain which route is currently best. In this non-stationary environment, one either chooses the best-experienced route so far (i.e., exploit) or tries a route that was inferior in the past but now may be superior (i.e., explore). Which actions a commuter should take in a series of choices is a non-trivial problem as optimal decision-making requires factoring in uncertainty about the state of the environment. An actor who excessively exploits will fail to notice when another action becomes superior. Conversely, an actor who excessively explores incurs an opportunity cost by frequently forgoing the high-payoff option.

Our focus is on the timing of exploratory choices. People should explore when they are uncertain about the state of the environment. *Reflective* belief-updates do this by incorporating predictions about unobserved changes in the environment. For example, a reflective belief-updater would increase their belief that an inferior route has improved as more time passes since the last observation because it becomes more likely that disruptive construction will have completed. In contrast, a *reflexive* belief-updater is only informed by direct observations of rewards and, therefore, does not fully utilize environmental structure to update beliefs and guide actions resulting in randomly timed exploratory choices.

This distinction closely echoes contemporary dual-system reinforcement learning (RL) approaches in which a reflexive, computationally parsimonious model-free controller competes for control of behavior with a reflective, model-based controller situated in prefrontal cortex (Daw et al., 2005). Previous work on exploration and exploitation indicates that *reflective* choice is resource intensive, perhaps relying on prefrontal systems (Badre et al., 2012, Otto et al., 2014). Correspondingly, populations that have reduced executive function, such as those experiencing depressive symptoms, are impaired in reflective decision making (Blanco et al., 2013), as are individuals under a secondary task load that exhausts limited cognitive resources (Otto et al., 2014).

Here, we test the hypothesis that reflective exploration is mediated by prefrontal systems by examining differences in the functional Val158Met polymorphism within the *COMT* gene (rs4680). The *COMT* gene modulates dopamine levels in prefrontal cortex with Met allele carriers having lower COMT enzyme activity and, therefore, higher dopamine levels, compared to Val/Val homozygotes (Gogos et al., 1998, Yavich et al., 2007, Kaenmaki et al., 2010). Val/Val homozygotes tend to perform worse on executive tasks and display increased frontal activation that may reflect inefficient processing compared to Met-carriers (Blasi et al., 2005, Winterer et al., 2006, Tan et al., 2007). Animal studies examining set-shifting behavior also indicate

the crucial role of PFC dopamine (Stefani and Moghaddam, 2006), which can be manipulated by COMT (Tunbridge et al., 2004). In humans, the *COMT* genotype predicts participants' ability to adapt behavior on a trial-by-trial basis (Frank et al., 2007), has been associated with performance on reversal learning tasks (Nolan et al., 2004), and has been linked to uncertainty-based exploration (Frank et al., 2009). But, the influence of the Val158Met polymorphism on cognitive function is debated, with some conflicting results. A recent meta-analysis concluded that there was little or no association between *COMT* genotype and scores on a set of standard cognitive tests (e.g. the Wisconsin Card Sorting task), though a reliable association was found between Met/Met genotype and higher IQ (Barnett, Scoriels, & Munafò, 2008).

It may be that *COMT* genotype has a more specific or subtle influence on cognition than is measured by many of the standard behavioral tests. Here we directly assessed the role of *COMT* variation in an exploratory decision-making task. We use computational models, related to reflective and reflexive exploration, to provide a clearer picture of the behavioral data. The main prediction is that Met carriers will explore reflectively, whereas Val/Val homozygotes will rely on simpler reflexive strategies.

One possibility is that the additional dopamine available for Met carriers functions more as a reserve rather than to facilitate cognitive function in general. We predict that Met carriers will be more resilient when cognitive resources are taxed under dual-task load.

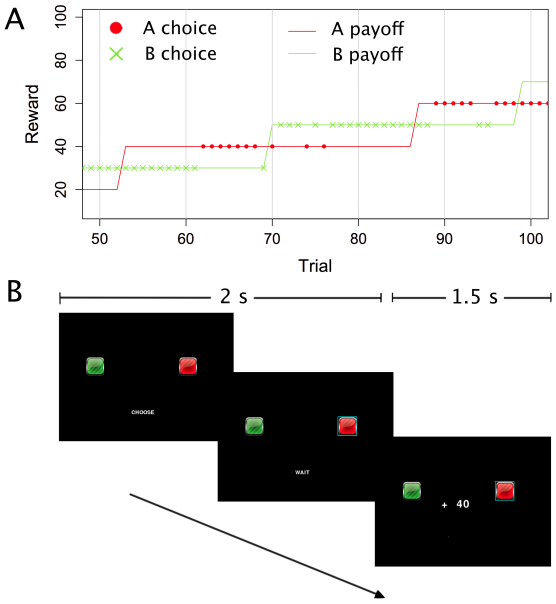

Figure 1: The Leapfrog task: example choices over 100 trials. On any trial the lower option might, with a probability of 0.075, increase its reward by 20 points, surpassing the other option. The relative superiority of the two options alternates as their reward values "leapfrog" over one another. The lines represent the true reward values, the dots a participant's choices.

2. Materials and Methods

We examined associations of *COMT* variants with exploratory strategies by using a paradigm termed the "Leapfrog" task (Knox et al., 2012), a variant of the "bandit" task (Sutton and Barto, 1998) that is specifically designed to evaluate exploratory behavior. In this task (Fig. 1), one of two options provides a higher reward than the other. With a fixed probability on each trial, the currently inferior option can increase in value, becoming the better option. Because the relative superiority of the options switches over time, participants must choose between *exploiting* the option with the highest observed reward and *exploring* to see whether the other option has surpassed it. This task is ideally suited to evaluate the timing of exploratory choices and to what extent they are guided by uncertainty in the environment, distinguishing reflective from reflexive choice strategies.

To tax mechanisms that support reflective exploration, which are thought to be resource intensive, participants in the dual-task condition also performed a tone counting task. Dual-task manipulations using tone counting are known to increase the prevalence of reflexive exploration strategies (Otto et al., 2014). More generally, secondary tasks that exhaust working memory resources tend to increase reliance on implicit strategies (Foerde et al., 2006, Zeithamova and Maddox, 2006) and cognitively inexpensive model-free choice strategies (Otto et al., 2013, Gershman et al., 2014).

2.1. *Models Evaluated*

We fit computational models that embody reflective and reflexive strategies to participants' data to evaluate their exploration behavior. The *Ideal Actor* model reflectively updates beliefs and plans ahead, taking into account the information gained by each choice and making choices that maximize long-term payoffs. Action-values are a product of both expected rewards and the potential to reduce uncertainty about the state of the environment. In contrast, the *Naïve RL* model instantiates the reflexive account of choice, in which the values of actions are based only on the rewards experienced so far. Its beliefs are updated reflexively in response to observed changes in rewards.

Both models incorporate a Softmax choice rule (Sutton and Barto, 1998), which chooses options as a function of the computed action-values. The Softmax inverse temperature is a free parameter in both models. Critically, the action-values used in the Softmax choice rule differ between the two models, leading to qualitative differences in exploratory behavior. The Naïve RL model explores with equal probability on every trial, whereas the probability of exploring increases after each successive exploitive choice for the Ideal Actor model (see Fig. 3A).

For the Naïve RL model, the value of each action is equal to the last observed reward for that action. The Ideal Actor computes action-values in two steps. First, it optimally updates its (Bayesian) beliefs about the state of the environment based on observations and its estimate of the environment volatility—a free parameter denoted $P(\text{flip})$. The Ideal Actor then optimally converts those beliefs into action-

values using established methods in RL (Kaelbling et al., 1996). More detailed descriptions of the models are provided in the Appendix. For full formal descriptions of these models, see Knox et al. (2012).

2.2. *Leapfrog Task*

Prior to the main task, participants passively viewed 500 training trials, in blocks of 100 trials. During these training trials, the rewards read CHANGED or SAME indicating whether the reward increased or not on that trial for each option. Each trial lasted for 0.5 s. Before each training block, participants estimated the number of jumps they expected to see in that block.

Participants then performed 300 trials of the main task, with a brief break after each 50 trial block. On each trial (see Fig. 1B) the word 'CHOOSE' appeared, and subjects had 1.5 s to select one of the two options by pressing a key on the keyboard. The chosen option was highlighted for the remainder of the trial. The reward received for the choice (e.g. '+ 60') was presented in the center of the screen for 1 s. Visual presentation of the reward was the only form of feedback. If the participant failed to respond in time the message 'TOO SLOW, TRY AGAIN' was displayed and the trial repeated. Every fourth time that they missed the response deadline, the experiment encouraged them to pay attention and respond more quickly. Immediately after reward presentation, the next trial began.

At the start of the main task, one option yielded a reward of 10 points and the other 20 points. On any trial the currently lower option could, with fixed probability of 0.075, increase by 20 points, becoming the higher-valued option. In this way, the two options alternate over time as the best option. Participants were informed of the initial starting values of the two options. They were also told that the two options would take turns being the better option and that the only way to know which was currently better was to sample the options. They were informed that the reward values options would change at the same rate that they observed in the training trials.

Participants were randomly assigned to either the single or dual task condition. In addition to the Leapfrog task, participants in the dual-task condition performed an auditory tone-counting task. On each trial, a series of high and low tones played and participants were instructed to count the total number of high tones over blocks of 50 trials while ignoring the low tones. At the end of each block, participants reported the total number of high tones in the block. At the start of each block, participants resumed counting from zero. The number of tones played per trial varied uniformly between 1 and 4. The base rate of high tones was determined every 50 trial block, selected randomly from a uniform distribution between .3 and .7. For example, for a base rate of .4, each tone played had a 40% chance of being a high tone. Tones occurred at random between 500 ms and 1750 ms after trial onset.

Participants were instructed that their goal was to earn as many points as possible during the task. They were told that there was a cash bonus associated with their

performance, and that the more points they earned the more they would be paid in cash. At the end of the experiment all participants were paid a $2 cash bonus. There were no specific instructions regarding incentive to perform well in the tone counting task.

2.3. *Participants*

Table 1: Demographic information by genotype. Standard deviations are in parentheses.

	Met/Met	Val/Met	Val/Val
N	44	94	71
Age, years	25.22 (4.96)	24.61 (4.26)	25.1 (4.26)
Gender	F = 31; M = 13	F = 47; M = 47	F = 41; M = 30
Years of education	15.2 (2.45)	15.32 (2.54)	15.8 (2.01)
Ethnicity			
Hispanic	10	19	12
Non-Hispanic	32	71	57
Decline to state	2	4	1
Race			
Asian	5	20	21
African American	1	4	3
Caucasian	34	53	40
Other	2	9	6
Decline to state	2	8	1

226 participants aged 18-35 were recruited from the greater Austin community through fliers and newspaper ads and received $10 (which includes the $2 bonus mentioned above) for their participation. Potential participants were screened for significant psychiatric disease via telephone using the Mini International Neuropsychiatric Interview (MINI), which screened for 17 different Axis I Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) disorders, including alcohol and drug abuse and/or dependence and attention-deficit hyperactivity disorder (ADHD). The MINI was chosen because of its acceptable validity, test-retest, and inter-rater reliability (Sheehan et al., 1998). Participants who met the MINI criteria for a current or past psychiatric diagnosis, currently taking psychoactive medication, currently in psychotherapy, or history of brain trauma were excluded from the study. Excluded participants were offered referrals to local mental health clinics.

Participants were excluded from further analysis when failing to satisfy one or more of three task engagement checks. (1) Choosing the lower value option (according to recent observations) for 10 or more consecutive trials (7 participants failed this check). (2) As explained in the Results section, being best fit by a baseline model of choice that was indicative of not processing sequential rewards (5 participants failed this check). (3) Having error of 70 or greater for two or more blocks in the secondary task of the dual-task condition (6 participants failed this check, including

one who also failed the first check). Overall, 17 participants were excluded, leaving 209 for further analysis. Checks 1 and 2 were designed to exclude participants that were not engaged in the Leapfrog task, while check 3 excludes participants that were not attending to the tone counting task. The Baseline model picks out participants that were not processing reward values for the entirety of the task, while check 1 picks out participants that failed to attend to reward values for a substantial portion of the task, repeatedly and consecutively choosing the option that was observably inferior.

2.4. *Genotyping*

For all participants genomic DNA was collected and isolated from buccal swabs using published procedures (Lench et al., 1988, Freeman et al., 1997). The *COMT* Val158Met polymorphism (rs4680) was genotyped using Taqman assay C_ _25746809_50 (Applied Biosystems) using an ABI 7900HT Real time PCR system. Our sample included 71 Val/Val homozygotes (mean age: 25.14; 41 female) and 138 Met carriers (mean age: 24.80; 78 female). Met carriers and Val homozygotes did not differ significantly in terms of age, $t(207)= 0.46$, **p**=0.65, gender, $\chi^2(1,N=209)=0.03$, **p**=.87, or years of education(15.28 vs. 15.80 years, respectively), $t(207)= 1.52$, **p**=0.13. This sample did not violate the Hardy-Weinberg equilibrium, $\chi^2(1,N=209)=1.52$, **p**=.22, indicating that the collected sample is representative of the population. In the results that follow, the qualitative pattern of results holds when the sample is restricted to Caucasians.

3. Results

Preliminary analyses evaluated the overall rate of exploration and participants' performance in the tone counting task in the dual-task condition. Subsequent analyses evaluate the main hypotheses by considering the sequencing of exploratory choices.

3.1. *Preliminary Analyses*

Met carriers and Val/Val homozygotes did not significantly differ in their mean error rate (21.85 vs. 17.51) in the tone counting task, $t(97)=1.76$, **p**=.08. Also easing the interpretation of the main analyses, Met carriers and Val/Val homozygotes did not significantly differ in their overall rate (.156 vs. .159) of exploration in the Leapfrog task across conditions, $t(207)=.36$, **p**=.72, nor within the single (.155 vs. .161) and dual (.156 vs. .156) task conditions, $t(108)=.60$, **p**=.55; $t(97)=.03$, **p**=.97, respectively. The two groups also did not differ in the mean number of flips that they predicted (8.49 vs. 8.79) in training blocks of 100 trials, $t(207) = 0.95$, **p**=0.34.

3.2. *Basic Measures of Performance in the Leapfrog Task*

Overall, the results support the main hypothesis that Met carriers will be more reflective in their exploration choices than Val/Val homozygotes when under dual-task conditions. This conclusion is supported by consideration of choice (Fig. 2a) and response time data (Fig. 2b).

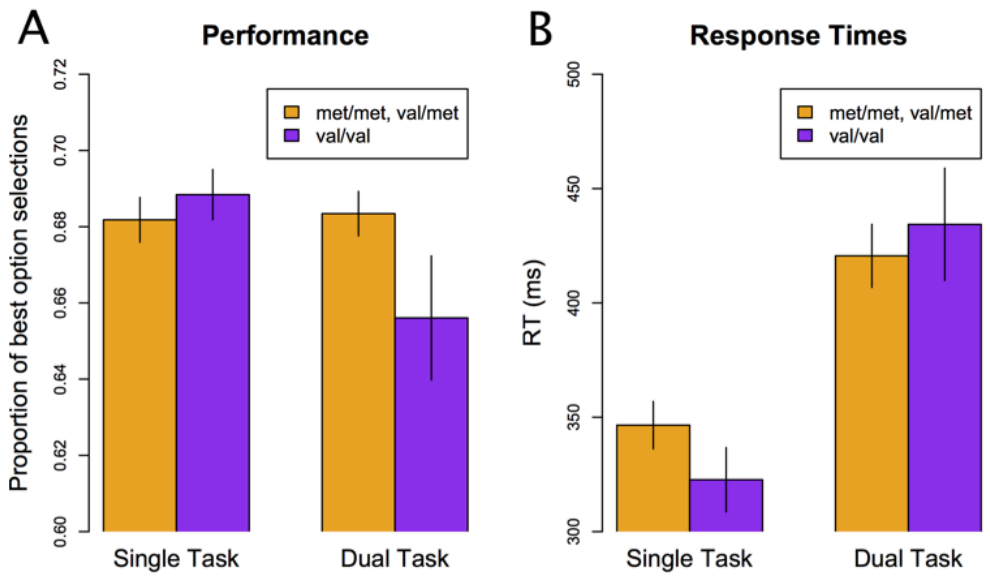

Figure 2: The main behavioral results. (a) Met Carriers (N = 138; 72 Single Task; 66 Dual Task) tended to choose the higher-valued option under dual-task conditions than did Val/Val homozygotes (N = 71; 38 Single Task; 33 Dual Task), indicated by a significant task X genotype interaction, $F(1,205)=4.17$, $p<.05$. (b) Response times for both groups and task conditions. Error bars reflect standard errors.

The choice data shown in Fig. 2a were analyzed using a task X genotype ANOVA. The main prediction was confirmed – there was an interaction such that Met carriers tended to perform better under dual-task conditions, $F(1,205)=4.17$, $\boldsymbol{p}<.05$. The main effects for task and genotype were not significant, $F(1,205)=1.54$, **p**=.22; $F(1,205)=1.26$, **p**=.26, respectively. While the two groups did not differ significantly in tone counting performance that Met carriers performed better on the Leapfrog task, but made slightly more errors in tone counting, suggests that these differences might result from Met carriers allocating more cognitive resources to the main task. To investigate this possibility we tested for a correlation between tone counting error rate and Leapfrog task performance across all dual-task participants. The correlation was not significant, r = -0.04, $t(97)$ = -0.39, **p**=0.70, providing evidence against this hypothesis.

Because heterozygote (Val/Met) performance appeared intermediate between Met homozygote and Val homozygote performance, and because some studies report gradual effects of *COMT* genotype (e.g Egan et al., 2001) we performed a task X genotype ANOVA with genotype coded as the number of Val alleles. There was a marginal task X genotype interaction on performance in this analysis, $F(1,205)=3.63$, **p**=0.058.

The response time data shown in Fig. 2b were also analyzed using a task X genotype ANOVA. The results were consistent with a speed-accuracy tradeoff in which participants were slower under more difficult conditions as participants were slower in the dual-task condition, $F(1,205)=18.88$, **p**<.05. The main effect of genotype was not significant, $F(1,205)=0.12$, **p**=.73, nor was the interaction, $F(1,205)=0.91$, **p**=.34.

Overall, these analyses converge in support of the main hypotheses. The dual-task manipulation exhausted limited cognitive resources and this adversely affected

performance in the Leapfrog task. As predicted, Met Carriers showed greater resilience under this load.

3.3. *Analyses of Reflective vs. Reflexive Choice*

The previous analyses considered basic performance measures. Although these analyses bear on the question of reflective vs. reflexive exploration, they did not directly evaluate participants' patterns of exploration. In this subsection, indices of reflective and reflexive exploratory choice are measured directly from the data and inferred using computational models to further evaluate the main hypotheses.

The Ideal Actor and Naive RL models were used to evaluate whether participants were reflective or reflexive explorers. The Ideal Actor's propensity to explore increases the longer it has been since an exploratory choice because uncertainty about which option is better grows, whereas the Naive RL model explores randomly independently of uncertainty (see Fig. 3a). We fit the Ideal Actor and the Naïve RL model to participants' trial-by-trial choice data by conducting an exhaustive grid search to find the set of parameters that maximized the likelihood of each model for each participant. Because the two models have different numbers of free parameters, we determined which model best characterized each participant using the Bayesian Information Criterion (BIC; Schwarz, 1978), mirroring previous used methods (Knox et al., 2012, Blanco et al., 2013). This same procedure was used to fit a baseline model mentioned in the Methods section – participants that were best characterized by a baseline model that assumed a fixed probability for selecting the left option (i.e., a model that is not reactive to the rewards in the task) were excluded from all analyses. Best-fitting parameter values for each group and condition are listed in Table 2 in the Appendix.

Each participant was classified as either an Ideal Actor or Naive RL model explorer depending on their BIC scores (see Fig. 3b). As predicted, there was a significant task X model interaction, G^2=6.40, **p**<.05, such that the Naive RL model better characterized participants under dual-task conditions. As predicted, there was a significant task X model X genotype interaction, G^2=3.89, **p** <.05, such that Val/Val homozygotes were disproportionately more likely to be characterized by the Naive RL model under dual-task conditions in comparison to Met carriers.

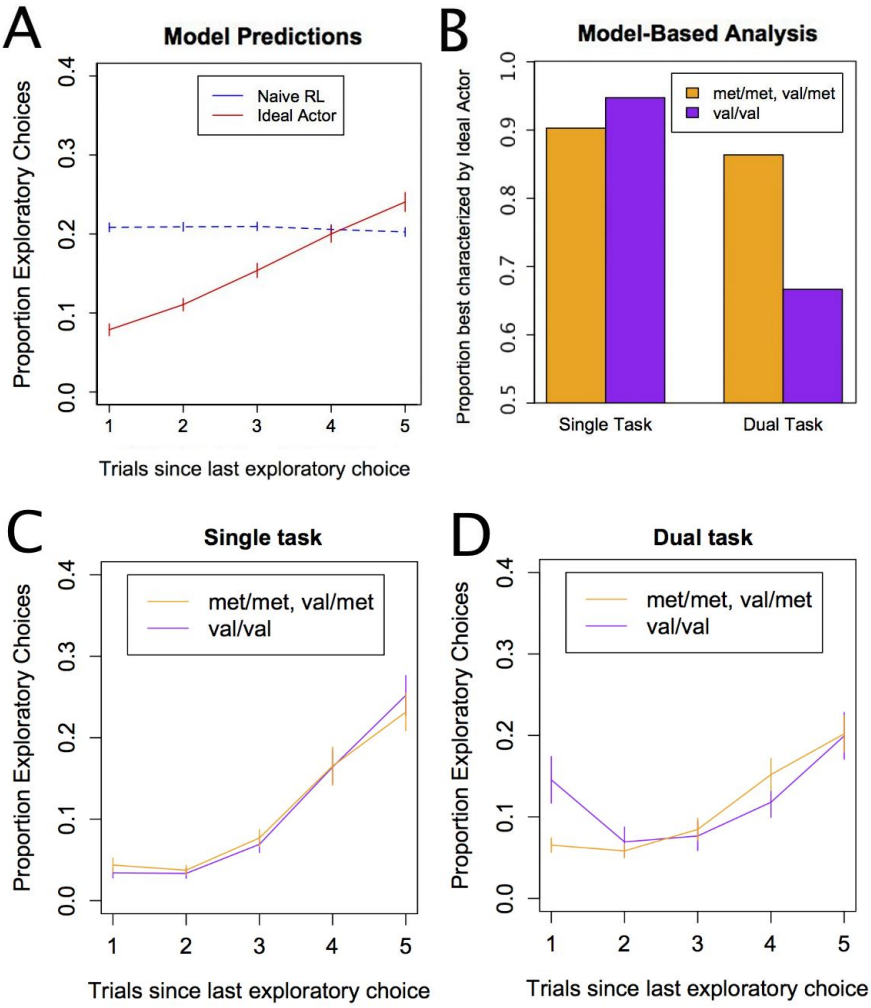

Figure 3: Exploration patterns characteristic of reflective and reflexive processing. (a) The Naive RL model predicts a constant probability of exploration as a function of the number of trials since the last exploration, whereas the Ideal Actor model is more likely to explore the longer it has been since the last exploration. (b) The Ideal Actor and Naive RL model were fit to each participant's data to determine which model best characterized each participant's choices. (c) Participants' choices in the single-task condition conform to the predictions of the reflective model. (d) In the dual-task condition the genotype groups show different patterns of exploratory choices. Error bars reflect standard errors.

These results support the hypothesis that Met carriers are more likely to remain reflective explorers when under cognitive load. As shown in Fig. 3A, the Naive RL and Ideal Actor models predict different functional relationships between the probability of exploring and the number of trials since an explorative choice. To confirm that these model signatures are present in the behavioral data, we evaluated whether participants displayed a flat slope (i.e., intercept only) relationship between the probability of exploration and the number of trials since the last exploratory choice as in the Naive RL model or a rising slope (i.e., linear with slope and intercept) as in the Ideal Actor model. BIC was again used to compare these two options to determine which regression model best characterized each participant's choices (see Fig. 3C). There was a significant task X model X genotype

interaction, G^2=6.91, **p**<.05, indicating that Met carriers were more likely to maintain the linear exploration pattern under dual-task conditions. There was also a significant task X model interaction, G^2=4.39, **p**<.05, indicating that the intercept-only model was more prevalent under dual-task conditions. These regression analyses parallel the full model-based analyses of the Naive RL and Ideal Actor models. We also directly analyzed exploration rates as a function of number of trials since the last exploratory choice (Fig. 3C and D) by genotype and task. A task X genotype X trial (since last exploratory choice) ANOVA revealed a significance 3-way interaction, $F(1,1036)=5.31$, **p**=0.02, which confirms that the genotype groups differ on this measure in the dual-task condition. As expected there were also a main effect of trial, $F(1,1036)= 219.49$, **p**<0.05, and a task X trial interaction, $F(1,1036)=16.4$, **p**<0.05*.* No other main effects or interactions reached significance.

One question is whether genotype predicts exploratory behavior above and beyond common measures of cognitive ability. As part of a larger data collection effort independent of this study's design, a number of measures, including WAIS vocabulary, logical memory immediate recall test, logical memory, digit span, and the Stroop task, were collected in a separate session. These measures, along with genotype, were entered into a stepwise regression to predict whether each participant in the dual-task condition followed the Ideal Actor. The stepwise regression solution only included digit span, z=1.61, **p**=.11, and genotype, z=1.97, **p**<.05, as predictors. This result indicates that genotype significantly predicts exploration behavior even when common measures of cognitive function are included as competing alternative predictors.

4. Discussion

Consistent with past research, participants' exploratory behavior was reflective under single-task conditions and became more reflexive under dual-task conditions that taxed limited cognitive resources. The main hypothesis tested was that Met carriers would be more resilient in the face of cognitive load and continue to explore reflectively. This hypothesis is based on how the *COMT* gene modulates dopamine levels in prefrontal cortex (Gogos et al., 1998, Yavich et al., 2007, Kaenmaki et al., 2010) and the associated performance differences between Met carriers and Val/Val homozygotes in cognitive tasks (Blasi et al., 2005, Winterer et al., 2006, Tan et al., 2007). Our analyses, which involved fits of computational models of reflective and reflexive exploration and standard statistical tests, converge in support of our a priori predictions.

An important aspect of our results is that the differences between genotypes only emerged under dual-task conditions. When the task became cognitively demanding and executive resources were taxed, Met carriers performed better than Val/Val homozygotes. This may be an important difference between out task and other cognitive tasks for which conflicting results have been found or which meta-analyses suggest may not be influenced by *COMT* genotype (Barnett, Scoriels, & Munafò, 2008). It may be that the higher levels of dopamine associated with having

the Met allele affords an individual a greater reserve of resources available when needed under taxing conditions, but which does not improve functioning under less demanding conditions. The level of cognitive demand of many standard cognitive tasks like the Wisconsin Card Sorting task may be more comparable to that of our single task condition.

These findings advance our understanding of the mechanisms of exploratory behavior and suggest that reflective exploratory behavior shares a neural basis with the frontal systems critical in model-based reinforcement learning (Daw et al., 2005, Badre et al., 2012). Frontal-dopamine systems appear crucial for maintaining, manipulating, and evaluating representations of the environment across a number of related tasks. One possible mechanism for this effect is that the Met allele has been associated with a greater prefrontal neuronal signal-to-noise ratio (Egan et al., 2001; Winterer et al., 2006). It could be that higher signal-to-noise ratio promotes greater stability in maintaining the representation of the environment, producing better estimates of current uncertainty and enabling more effective reflective updating of the environmental representation. Better representations of the environment allow directed exploratory choices, like those produced by our Ideal Actor model that optimally maintains and updates the environmental representation. Without accurate representations, exploratory choices become less structured, as in our Naïve RL model.

In addition to elucidating the cognitive and neural mechanism underlying exploratory behavior, we hope our tasks and analysis methods will prove useful to researchers considering related questions. For example, our Leapfrog task and modeling approach are readily adapted to non-human animal studies, which presents exciting possibilities such as manipulating *COMT* expression directly. Some limitations of this contribution include relatively low sample size and the possibility of a third variable explanation including but not limited to another polymorphism in linkage disequilibrium with rs4680 or within-ethnicity population stratification. Additionally, our study lacks a replication sample. An independent replication of this finding to determine the robustness of the effect is an important direction for future research. Nevertheless, given the broad literature describing the cognitive consequences of variation in the *COMT* gene, and our attempts to reduce the risk of population stratification by conducting ethnicity-specific analyses, we believe these data add substantively to the literature and provide direction for future studies.

References

Badre D, Doll BB, Long NM, Frank MJ (2012) Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron 73:595-607.

Barnett JH, Scoriels L, Munafò MR (2008) Meta-analysis of the cognitive effects of the catechol-O-methyltransferase gene Val158/108Met polymorphism. Biological Psychiatry 64:137-144.

Blanco NJ, Otto AR, Maddox WT, Beevers CG, Love BC (2013) The influence of depression symptoms on exploratory decision-making. Cognition 129:563-568.

Blasi G, Mattay VS, Bertolino A, Elvevag B, Callicott JH, Das S, Kolachana BS, Egan MF, Goldberg TE, Weinberger DR (2005) Effect of catechol-O-methyltransferase val158met genotype on attentional control. J Neurosci 25:5038-5045.

Cassandra A, Littman ML, Zhang NL (1997) Incremental Pruning: A Simple, Fast, Exact Method for Partially Observable Markov Decision Processes. In Proceedings Of The Thirteenth Conference On Uncertainty In Artificial Intelligence. 54–61.

Cohen JD, McClure SM, Yu AJ (2007) Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical transactions of the Royal Society of London Series B, Biological sciences 362:933-942.

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704-1711.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876-879.

Egan MF, Goldberg TE, Kolachana BS, Callicott JH, Mazzanti CM, Straub RE, Goldman D, Weinberger DR. (2001) Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. Proc Natl Acad Sci U S A. 2001 Jun5;98(12):6917-22. Epub 2001 May 29. PubMed PMID: 11381111; PubMed Central PMCID: PMC34453.

Foerde K, Knowlton BJ, Poldrack RA (2006) Modulation of competing memory systems by distraction. Proceedings of the National Academy of Sciences 103:11778-11783.

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci 12:1062-1068.

Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE (2007) Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A 104:16311-16316.

Freeman B, Powell J, Ball D, Hill L, Craig I, Plomin R (1997) DNA by mail: an inexpensive and noninvasive method for collecting DNA samples from widely dispersed populations. Behavior genetics 27:251-257.

Gershman SJ, Markman AB, Otto AR (2014) Retrospective revaluation in sequential decision making: a tale of two systems. J Exp Psychol Gen 143:182-194.

Gogos JA, Morgan M, Luine V, Santha M, Ogawa S, Pfaff D, Karayiorgou M (1998) Catechol-O-methyltransferase-deficient mice exhibit sexually dimorphic changes in catecholamine levels and behavior. Proc Natl Acad Sci U S A 95:9991-9996.

Hills TT, Todd PM, Lazer D, Redish AD, Couzin ID, Cognitive Search Research G (2015) Exploration versus exploitation in space, mind, and society. Trends Cogn Sci 19:46-54.

Kaelbling LP, Littman ML, Moore AP (1996) Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research 4:237-285.

Kaenmaki M, Tammimaki A, Myohanen T, Pakarinen K, Amberg C, Karayiorgou M, Gogos JA, Mannisto PT (2010) Quantitative role of COMT in dopamine clearance in the prefrontal cortex of freely moving mice. Journal of neurochemistry 114:1745-1755.

Knox WB, Otto AR, Stone P, Love BC (2012) The nature of belief-directed exploratory choice in human decision-making. Front Psychol 2:398.

Lench N, Stanier P, Williamson R (1988) Simple non-invasive method to obtain DNA for gene analysis. Lancet 1:1356-1358.

Nolan, K. A., Bilder, R. M., Lachman, H. M., & Volavka, J. (2004). Catechol O-methyltransferase Val158Met polymorphism in schizophrenia: differential effects of Val and Met alleles on cognitive stability and flexibility. *American Journal of Psychiatry*, *161*(2), 359-361.

Otto AR, Gershman SJ, Markman AB, Daw ND (2013) The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psychol Sci 24:751-761.

Otto AR, Knox WB, Markman AB, Love BC (2014) Physiological and behavioral signatures of reflective exploratory choice. Cogn Affect Behav Neurosci 14:1167-1183.

Schwarz GE (1978) Estimating the dimension of a model. Annals of Statistics 6:461–464.

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, Hergueta T, Baker R, Dunbar GC (1998) The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. The Journal of clinical psychiatry 59 Suppl 20:22-33;quiz 34-57.

Stefani MR, Moghaddam B (2006) Rule learning and reward contingency are associated with dissociable patterns of dopamine activation in the rat prefrontal cortex, nucleus accumbens, and dorsal striatum. J Neurosci 26:8810-8818.

Sutton RS, Barto AG (1998) Reinforcement learning: An Introduction. Cambridge, MA: MIT Press.

Tan HY, Chen Q, Goldberg TE, Mattay VS, Meyer-Lindenberg A, Weinberger DR, Callicott JH (2007) Catechol-O-methyltransferase Val158Met modulation of prefrontal-parietal-striatal brain systems during arithmetic and temporal transformations in working memory. J Neurosci 27:13393-13401.

Tunbridge EM, Bannerman DM, Sharp T, Harrison PJ (2004) Catechol-o-methyltransferase inhibition improves set-shifting performance and elevates

stimulated dopamine release in the rat prefrontal cortex. J Neurosci 24:5331-5335.

Winterer G, Musso F, Vucurevic G, Stoeter P, Konrad A, Seker B, Gallinat J, Dahmen N, Weinberger DR (2006) COMT genotype predicts BOLD signal and noise characteristics in prefrontal circuits. Neuroimage 32:1722-1732.

Yavich L, Forsberg MM, Karayiorgou M, Gogos JA, Mannisto PT (2007) Site-specific role of catechol-O-methyltransferase in dopamine overflow within prefrontal cortex and dorsal striatum. J Neurosci 27:10196-10209.

Zeithamova D, Maddox WT (2006) Dual-task interference in perceptual category learning. Mem Cognit 34:387-398.

Appendix: Model details

Naïve Reinforcement Learning (RL) Model

The Naïve RL model reflexively updates its beliefs about reward values based only on its observations (i.e. it believes that the payoffs for each option are as they were last observed). In brief, the model simply remembers which option was best based on recent observations and exploits that option with a fixed probability, exploring the remainder of the time.

Accordingly, the model assumes that Action *H* (that with highest observed reward) and ¬*H* (the alternative action with lower observed reward) give rewards of 1 and 0 respectively, corresponding to the higher and lower payoffs. Its expectation of each action's reward, $Q(H)$, is input into a Softmax choice rule (Sutton & Barto, 1998), resulting in a constant probability of exploring or exploiting:

$$P(H)=\exp(\gamma \cdot Q(H))/[\exp(\gamma \cdot Q(H))+\exp(\gamma \cdot Q(\neg H))]$$

where γ is an inverse temperature parameter, referred to as the Softmax parameter in the text. As γ increases, the probability that the highest-observed action (H) will be chosen increases. As γ approaches zero, the model moves towards choosing actions with uniform randomness. This parameter is the only free parameter in the model.

Ideal Actor Model

The reflective Ideal Actor model maintains optimal beliefs about the probability that each option will give a higher immediate reward. These beliefs are then used to compute optimal action values. The model has two free parameters: $P(\text{flip})$, its estimate of how often flips occur, which it uses to perform belief updates, and γ, the inverse temperature parameter used in the Softmax choice rule, as in the Naïve RL model above.

The Ideal Actor model maintains a probabilistic distribution over possible underlying environment states, represented as a belief *B*, which is the probability that the exploitative action—i.e., choosing the option with the currently highest observed reward—will actually yield the larger immediate reward. The underlying environment state can be formulated as the number of unobserved (i.e., true) flips at a given time point. If there are 0 or 2 unobserved flips, then the exploitative action will yield the higher reward. If there is 1 unobserved flip, then the option with the lower observed reward yields the true higher immediate reward. Beliefs are optimally updated following each choice and observation of the resulting reward. Updating the belief B_{t+1}—the probability distribution over the number of unobserved flips (0, 1, or 2) before taking the action at trial *t+1*—depends on the previous belief state B_t, the action taken at trial *t* (exploratory or exploitative), the observed number of flips seen as a result of that action and the assumed volatility

rate of the environment—the free parameter *P*(flip). Individual state transition probabilities based on these factors are combined and normalized to form a posterior belief:

$$B_{t+1} = \frac{P(s_{0,t+1}, s_{0,t}) + P(s_{2,t+1}, s_{1,t})}{P(s_{0,t+1}, s_{0,t}) + P(s_{2,t+1}, s_{1,t}) + P(s_{1,t+1}, s_{1,t}) + P(s_{1,t+1}, s_{0,t})}$$

where the state $s_{i,t+1}$ refers to the number of unobserved flips after the choice and reward observation were made while $s_{i,t}$ refers to the number of unobserved flips before the choice.

Using these optimally maintained beliefs, the Ideal Actor then optimally computes action-values for each action. This is accomplished by formulating the task as a Partially Observable Markov Decision Process (POMDPs) and using methods for solving POMDPs. To calculate these optimal action values, we employed Cassandra et al.'s (Cassandra, Littman, & Zhang, 1997) incremental pruning algorithm, an exact inference method that calculates values for each possible belief state at each time horizon (i.e., number of choices remaining). These routines are implemented in Cassandra et al.'s POMDP-Solve library (Cassandra et al., 1997). The resulting action values express the statistical expectation of the sum of all future reward given that the option is chosen and assuming that all subsequent choices are performed optimally. The true Ideal Actor deterministically chooses the option with the highest resulting action value. However, for the purpose of fitting the model to participants' choice data, we use a Softmax choice rule (identical to that used by the Naïve RL model) to generate response probabilities from these action values. The Softmax inverse temperature γ is a free parameter.

Table 2. Best-fitting parameter values for each condition and group for the reflective Ideal Actor and the reflexive Naïve Reinforcement Learning models. Standard errors of the mean are listed in parentheses.

	Ideal Actor		**Naïve RL model**
	P(flip)	Softmax parameter	Softmax parameter
Single-Task			
Met Carriers	0.032 (0.005)	0.163 (0.007)	0.148 (0.003)
Val Homozygotes	0.036 (0.005)	0.160 (0.006)	0.146 (0.004)
Dual-Task			
Met Carriers	0.033 (0.007)	0.152 (0.005)	0.149 (0.004)
Val Homozygotes	0.016 (0.004)	0.127 (0.009)	0.137 (0.006)