HHS Public Access

Author manuscript

*J Cogn Neurosci*. Author manuscript; available in PMC 2015 May 01.

Published in final edited form as:

*J Cogn Neurosci*. 2015 February ; 27(2): 319–333. doi:10.1162/jocn_a_00709.

# Cognitive Control Predicts Use of Model-Based Reinforcement-Learning

**A. Ross Otto**[*],
Center for Neural Science, New York University

**Anya Skatova**[*],
School of Psychology and Horizon Digital Economy Research Institute, University of Nottingham

**Seth Madlon-Kay**, and
Duke Institute for Brain Sciences, Duke University

**Nathaniel D. Daw**
Center for Neural Science and Department of Psychology, New York University

## Abstract

Accounts of decision-making and its neural substrates have long posited the operation of separate, competing valuation systems in the control of choice behavior. Recent theoretical and experimental work suggest that this classic distinction between behaviorally and neurally dissociable systems for habitual and goal-directed (or more generally, automatic and controlled) choice may arise from two computational strategies for reinforcement learning (RL), called model-free and model-based RL, but the cognitive or computational processes by which one system may dominate over the other in the control of behavior is a matter of ongoing investigation. To elucidate this question, we leverage the theoretical framework of cognitive control, demonstrating that individual differences in utilization of goal-related contextual information—in the service of overcoming habitual, stimulus-driven responses—in established cognitive control paradigms predict model-based behavior in a separate, sequential choice task. The behavioral correspondence between cognitive control and model-based RL compellingly suggests that a common set of processes may underpin the two behaviors. In particular, computational mechanisms originally proposed to underlie controlled behavior may be applicable to understanding the interactions between model-based and model-free choice behavior.

## Introduction

A number of theories across neuroscience, cognitive psychology, and economics posit that choices may arise from at least two distinct systems (Balleine & O'Doherty, 2009; Daw, Niv, & Dayan, 2005; Dolan & Dayan, 2013; Kahneman, 2011; Loewenstein, 1996). A recurring theme across these dual-system accounts is that the systems rely differentially upon automatic or habitual versus deliberative or goal-directed modes of processing.

Please address all correspondence to: A. Ross Otto, Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003, rotto@nyu.edu.

[*]The first two authors contributed equally to this article, and ordering was determined arbitrarily.

A popular computational refinement of this idea, derived initially from computational neuroscience and animal behavior, proposes that the two modes of choice arise from distinct strategies for learning the values of different actions, which operate in parallel (Daw et al., 2005). In this theory, habitual choices are produced by model-free reinforcement learning (RL), which learns which actions tend to be followed by rewards. This is the approach taken by prominent computational models of the dopamine system (Schultz, Dayan, & Montague, 1997). In contrast, goal-directed choice is formalized by model-based RL, which reasons prospectively about the value of candidate actions using knowledge (a learned internal "model") about the environment's structure and the organism's current goals. Whereas model-free choice involves requires merely retrieving the (directly learned) values of previous actions, model-based valuation is typically envisioned as requiring a sort of mental simulation – carried out at decision time – of the likely consequences of candidate actions, using the learned internal model. Informed by these characterizations, recent work reveals that under normal circumstances, reward learning by humans exhibits contributions of both putative systems (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gläscher, Daw, Dayan, & O'Doherty, 2010), and these influences are behaviorally and neurally dissociable.

Under this framework, at any given moment both the model-based and model-free systems can provide action values to guide choices, inviting a critical question: how does the brain determine which system's preferences ultimately control behavior? Despite progress characterizing each system individually, little is yet known about how these two systems interact, such as how the brain arbitrates between each system's separately learned action values. How these two systems jointly influence behavior is important, in part, because disorders of compulsion such as substance abuse have been argued to stem from an imbalance in expression of the two systems' values, favoring the more habitual, model-free influences (Everitt & Robbins, 2005; Kahneman, 2011; Voon et al., in press).

A separate research tradition, grounded in neuropsychiatry and human cognitive neuroscience, has investigated a similar question: how individuals hold in mind contextual, task-related information in order to flexibly adapt behavior and direct cognitive processing in accordance with internally maintained goals. One key example of this sort of cognitive control is the ability for internally maintained goals to overcome prepotent and/or stimulus-driven responses, as most famously operationalized in the classic Stroop task (Cohen, Barch, Carter, & Servan-Schreiber, 1999). This work has stemmed a rich set of experiments and models describing the brain's mechanisms for cognitive control (Braver, 2012).

Considering these two traditionally separate lines of work together yields a compelling, but underexplored conceptual similarity: cognitive control and model-based RL both characteristically entail leveraging higher-order representations in order to overcome habitual, stimulus-driven actions (Braver, 2012). In particular, we hypothesize that model-free action tendencies are analogous to prepotent color reading in Stroop, and the ability to instead act in accord with the evaluations of a model-based system is equivalent to biasing behavior toward higher-level representations of goals and context—in the RL case, the representation of the internal model. Interestingly, the two literatures have a complementary relationship in terms of their key questions. Historically, the RL work speaks little to how higher-order representations interact with the simpler ones to influence choice, and instead

focuses on how each system learns and computes action values. The cognitive control literature, on the other hand, contains much research on how contextual information is used to override prepotent actions but concentrates less (though see Collins and Frank, 2013) on how these different competing representations are learned in the first place.

Highlighting this thematic correspondence, theoretical accounts of cognitive control and goal-directed choice posit complementary or even overlapping computational mechanisms (Alexander & Brown, 2011; Botvinick, Braver, Barch, Carter, & Cohen, 2001; Daw et al., 2005) and neural data on either function suggest the involvement of nearby (or overlapping) prefrontal structures (Alexander & Brown, 2011; Economides, Guitart-Masip, Kurth-Nelson, & Dolan, 2014; Holroyd & Yeung, 2012; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011). Here we hypothesize that the same mechanisms characterized by well-known cognitive control tasks may also support the expression of model-based, over model-free behaviors in sequential choice. One previous result supporting this notion comes from the category learning literature, wherein older adults' Stroop and set-shifting performance was predictive of task-appropriate, rule-based learning (which may be analogous to model-based RL) over another, incremental learning strategy (which may be analogous to model-free RL; Maddox et al. 2010).

Across two experiments, we test the operational relationship between the two constructs. Specifically, we examine how individual differences in cognitive control, assessed using two separate paradigms, predict the behavioral contributions of model-based RL in a separate, sequential choice task. Indeed, stable individual differences in cognitive control ability are thought in part to reflect differences in controlled or executive-dependent processing (Kane & Engle, 2003). If the two functions share common underpinnings, then we should expect their behavioral expression in disparate tasks to correlate.

## Methods

In each of two experiments, participants completed two tasks: a test of cognitive control and a sequential choice task in which the behavioral contributions of model-based versus model-free learning to can be independently assessed (Daw et al., 2011). We then examined the relationship between individual differences in behavior across the two tasks.

### Experiment 1

**Participants—**47 participants undertook two behavioral tasks: a Stroop task and a sequential decision-making task. These participants constituted a subsample of the data of Skatova et al. (2013) where other aspects of their behavior were reported. Two participants exhibited an error rate greater than 25% on the Stroop task and their data were excluded, leaving 45 participants for the reported results. All participants gave written consent prior to the study, and were paid a fixed amount plus a bonus contingent on their decision task performance. The study was approved by NYU's UCAIHS.

**The Stroop task—**Participants performed a computerized version of the Stroop task (Besner, Stolz, & Boutilier, 1997), which required them to identify, as quickly and as accurately as possible, in which one of three colors the word on the screen was presented. In

each trial, prior to the stimulus, participants saw a fixation cross in the center of the screen for 200ms. One of three color words (“RED”, “GREEN,” or “BLUE”) was displayed either in red, green or blue on a black background, in 20-point Helvetica bold font, until participants responded. Participants responded using labeled keys on the keyboard (“b” = blue, “g” = green, “r” = red). There was an inter-trial interval of 250ms.

Participants received two blocks, each of 120 trials. In one block, (“incongruent infrequent”), 80% of the trials were congruent (e.g. RED in red type) while 20% were incongruent (e.g. RED in blue type), whereas in the other block type (“incongruent frequent”) these proportions were reversed. The order of blocks was counterbalanced across participants. Participants were not informed about the differences between the blocks and were given a short break between blocks. Prior to each experimental block, participants received a block of 24 practice trials. In the practice trials, but not the experimental trials, participants received feedback on the screen indicating whether their response was correct or not. Trials in which participants made errors (averaging 2.1%) were excluded from analysis.

**Two-Step Decision-Making Task—**Each participant undertook 350 trials of the two-stage decision task (Figure 1A) described in detail by Daw et al. (2011). On each trial, an initial choice between two options labeled by Tibetan characters led probabilistically to either of two, second-stage “states,” represented by different colors. Each first stage choice was associated with one of the second stage states, and led there 70% of the time. In turn, each of the second-stage states demanded another choice between another pair of options labeled by Tibetan characters. Each second-stage option was associated with a different probability of delivering a monetary reward (versus nothing) when chosen. To encourage participants to continue learning throughout the task, the chances of payoff associated with the four second-stage options were changed slowly and independently throughout the task, according to Gaussian random walks. In each stage participants had 2s to make a choice. Inter-stimuli and inter-trial intervals were 500ms and 300ms, respectively, and monetary reward was presented for 500ms.

**Data analysis—**For each subject, we first estimated their Stroop Incongruency Effect (IE) on correct trials for incongruent trials in each block type using a linear regression with RTs as the outcome variable and explanatory variables that crossed the block type (incongruent infrequent versus frequent) with indicators for congruent and incongruent trials. Before being entered into the regression, RTs were first log-transformed to remove skew (Ratcliff, 1993) and then z-transformed with respect to RTs on all correct trials. Further, the linear model contained an additional nuisance variable to remove the influence of stimulus repetitions, documented to facilitate faster RTs (Kerns et al., 2004). Each participant’s individual regression yielded two coefficients of interest: the IE for the incongruent-infrequent block and the IE for the incongruent-frequent block.

To assess model-based and model-free contributions to trial-by-trial learning, we conducted a mixed-effects logistic regression to explain participants’ first stage choices on each trial (coded as stay or switch relative to previous trial) as a function of the previous trial’s outcomes (whether or not a reward was received on the previous trial and whether the previous transition was common or rare). Within-subject factors (the intercept, main effects

of reward and transition, and their interaction) were taken as random effects across subjects, and estimates and statistics reported are at the population level. To assess whether these learning effects covaried with the Stroop effect, the four variables above were each additionally interacted, across subjects, with the infrequent IE and frequent IE, entered into the regression as z-scores. The full model specification and coefficient estimates are provided in Table 2.

The mixed-effects logistic regressions were performed using the lme4 package (Pinheiro & Bates, 2000) in the R programming language. Linear contrasts were computed in R using the "esticon" function in the doBy package (Højsgaard & Halekoh, 2009). The individual model-based effects plotted in Figures 2A and B are the estimated per-subject regression coefficients from the group analysis (conditioned on the group level estimates) superimposed on the estimated group-level effect.

## Experiment 2

**Participants—**We recruited 83 individuals on Amazon Mechanical Turk, an online crowd-sourcing tool, to perform the Dot Pattern Expectancy task (DPX; MacDonald, 2008) followed by the sequential choice task. Mechanical Turk (www.mturk.com) allows users to post small jobs to be performed anonymously by "workers" for a small amount of compensation (Crump, McDonnell, & Gureckis, 2013; McDonnell et al., 2012). Participants were all US residents, paid a fixed amount ($2 USD) plus a bonus contingent on their decision task performance, ranging from $0–1 USD.

As internet behavioral data collection characteristically entails a proportion of participants failing to engage fully with the task—and, in particular, as in pilot studies using other tasks we noted a tendency toward lapses in responding and in the computer's focus on our task's browser window, which some queried participants indicated was due to "multitasking" (Boureau and Daw, unpublished observations)—for this study, we employed principled a priori criteria for participant inclusion in our analyses of interest, following recent recommendations (Crump et al., 2013). First, we excluded the data of 17 participants who missed more than 10 response deadlines in the DPX task and/or 20 deadlines in the two-stage task. Next, to ensure that the control measures yielded by the DPX task reflect understanding of the task instructions and engagement with the task, we excluded the data of 13 subjects who exhibited a d' of less than 2.5 in classifying target versus non-target responses. Critically, $d'$ is agnostic to performance on specific non-target trials (that is, a low $d'$ could arise from incorrect target responses during AY, BX, and/or BY trials) and therefore the engagement metric is not biased towards either of our putative measures of control. Finally, following our previous work (Otto, Gershman, Markman, & Daw, 2013), we employed a further step to remove participants who failed to demonstrate sensitivity to rewards in the decision task. In particular, using second-stage choices (a separate measure from the first-stage choices that are our main dependent measure), we excluded the data of 2 participants who repeated previously rewarded second-stage responses—i.e., P(stay|win)—at a rate less than 50%. A total of 51 subjects remained in our subsequent analyses.

**DPX Task—**Subjects first performed the DPX task. Each trial, a cue stimulus (either of two blue dot patterns, which we refer to as A or B) appeared for 500ms, followed by a fixation point for 2000ms (the delay period), followed by a probe stimulus for 500ms (either of two white dot patterns, labeled X or Y), followed by blank screen in which participants had 1000ms to make a target or non-target response using the "1" and "2" keys respectively. Subjects were instructed that the target cue-probe pair was AX, and all other sequences (AY, BX, BY) were non-targets. Feedback ("CORRECT," "INCORRECT", or "TOO SLOW") was provided for 1000ms, and the next trial began immediately. Following past work that sought to optimize the psychometric properties of the task to yield maximum sensitivity to individual differences (Henderson et al., 2012), participants completed 4 blocks of 32 trials consisting of 68.75% AX trials and 12.5% for each of the remaining trial types (AY, BX, and BY). The predominance of AX trials ensures that the target response is a prepotent response.

Immediately after the DPX task, participants were given choice task instructions and completed 10 practice trials to familiarize themselves with the two-stage task structure and response procedure, at which point they completed 200 trials of the full task. The two-stage task in Experiment 2 was the same as in Experiment 1, excepting three changes. First, the Tibetan characters used to represent the options at each stage were replaced with fractal images. Second, the inter-trial and inter-stimulus intervals were each set to 1000ms. Third, participants completed 200 trials, following Daw et al. (2011), rather than 350.

**Data Analysis—**To measure the influence cue-triggered contextual information in the DPX task, we considered individual differences in RT slowing for two non-target (and non-prepotent) trial types of interest (AY and BX). Following Experiment 1, we first applied a log-transformation of RTs to remove skew. We then standardized (i.e., z-scored) these RTs within each participant to equate participants with respect to global response speed and variance (Chatham, Frank, & Munakata, 2009; Paxton, Barch, Racine, & Braver, 2008). We then measured cognitive control according to the mean (z-scored) RT for each of two non-target trial types of interest. This analysis parallels the regression-based approach for assessing RTs in Experiment 1, the only difference being the inclusion of an additional explanatory variable factoring out stimulus repetition effects in the Stroop task, which were not a significant factor in the DPX. Response slowing on AY trials and speeding on BX trials, relative to all trials, is interpreted as reflecting control (Braver, Satpute, Rush, Racine, & Barch, 2005). We also assessed accuracy by calculating $d'$-*context* for each subject, which is robust to overall bias towards making target or non-target responses (Henderson et al., 2012). This quantity indexes task sensitivity (i.e., the traditional $d'$ from signal detection theory) using the hit rate from AX trials and the false alarm rate from BX trials.

Across three logistic regression models of RL task behavior (see Results section), we examined the interaction between either or both of these between-subjects control measures and the within-subject trial-by-trial learning variables (previous reward and transition type). The RT measures were entered into the regressions as z-scores, and the within-subject coefficients were taken as random effects over subjects. Individual model-based effect sizes (Figures 4A and B) were calculated from the respective reduced models (Tables 4 and 5), in

the same manner as in Experiment 1; the individual contrast reported in Figure 4C is calculated from linear contrasts taken on the regression model (Table 3).

## Results

### Experiment 1

Here we demonstrate that interference effects in the Stroop color-naming task relate to the expression of model-based choice in sequential choice. We first measured participants' susceptibility to interference in a version of the Stroop task (Kerns et al., 2004), in which subjects respond to the ink color of a color word (e.g., "RED") while ignoring its semantic meaning. In the Stroop task, cognitive control facilitates biasing of attentional allocation—strengthening attention to the task relevant feature and/or inhibiting task-irrelevant features—which in turn permits the overriding of inappropriate, prepotent responses (Braver & Barch, 2002). Of key interest was the incongruency effect (IE): the additional time required to produce a correct response on incongruent ("RED" in blue type) compared to congruent ("RED" in red type) trials. Incongruent trials require inhibition of the prepotent color-reading response, thought to be a critical reflection of cognitive control. We then examined whether an individual's IE predicted the expression of model-based strategies in sequential choice.

Forty-five participants completed two blocks of the Stroop task, with infrequent (20%) or frequent (80%) incongruent trials. We included the frequent condition as a control, because frequent incongruent trials should allow for stimulus-response learning (Bugg, Jacoby, & Toth, 2008; Jacoby, Lindsay, & Hessels, 2003) and/or adjustments in strategy or global vigilance (Bugg, McDaniel, Scullin, & Braver, 2011; Carter et al., 2000) to lessen the differential reliance on control in incongruent trials[1].

There were significant IEs in both the infrequent blocks ($M$=99.89ms, $t$=7.41, $p$<.001) and frequent blocks ($M$=61.76ms, $t$=7.63, $p$<.001), and, following previous work employing the proportion congruent manipulation (Carter et al., 2000; Lindsay & Jacoby, 1994; Logan & Zbrodoff, 1979) the IE was significantly larger in the infrequent compared to the frequent condition ($t$=2.43 $p$<.05). Table 1 reports the full pattern of median RTs and accuracies across conditions.

To identify contributions of model-based and model-free choice, participants subsequently completed a two-stage RL task (Daw et al., 2011) (Figure 1). In each two-stage trial, participants made an initial first-stage choice between two options (depicted as Tibetan characters), which probabilistically leads to one of two second-stage "states" (colored green or blue). In each of these states participants make another choice between two options, which were associated with different probabilities of monetary reward. One of the first-stage responses usually led to a particular second-stage state (70% of the time) but sometimes led to the other second-stage state (30% of the time). Because the second-stage reward

[1]Note that under a different interpretation, the relationship between the frequent and infrequent conditions could have the opposite directionality, since global strategic adjustments – to the extent they drive performance in the frequent condition and are most robustly exercised there – themselves may constitute a particular, proactive form of cognitive control (Bugg et al., 2011; Carter et al., 2000).

probabilities independently change over time, participants need to make trial-by-trial adjustments to their choice behavior in order to effectively maximize payoffs.

Model-based and model-free strategies make qualitatively different predictions about how second-stage rewards influence subsequent first-stage choices. For example, consider a first-stage choice that results in a rare transition to a second stage, wherein that second-stage choice was rewarded. A pure model-free strategy would predict repeating the same first-stage response because it ultimately resulted in reward. The predisposition to repeat previously reinforced actions in ignorance of transition structure is, in the dual-systems RL framework adopted here (Daw et al., 2005), a characteristic behavior of habitual control. A model-based choice strategy, utilizing a model of the environment's transition structure and immediate rewards to prospectively evaluate the first-stage actions, would predict a decreased tendency to repeat the same first-stage option because the other first-stage action was actually more likely to lead to that second-stage state.

These patterns by which choices depend on the previous trial's events can be distinguished by a two-factor analysis of the effect of the previous trial's reward (rewarded versus unrewarded) and transition type (common versus rare) on the current first-stage choice. The predicted choice patterns for a purely model-based and a purely model-free strategy are depicted in Figures 1B and C, respectively. A purely model-free strategy predicts that only reward should impact whether or not a first-stage action is repeated (a main effect of the reward factor) while a model-based strategy predicts that this effect of reward depends on the transition type, leading to a characteristic interaction of the reward effect by the transition type. In previous studies, human choices—and analogous effects on choice related BOLD signals—exhibit a mixture of both effects (Daw et al., 2011).

Following Daw et al. (2011), we factorially examined the impact of both the transition type (common versus rare) and reward (rewarded versus not rewarded) on the previous trial upon participants' tendency to repeat the same first-stage choice on the current choice. Consistent with previous studies (Daw et al., 2011; Otto, Raio, Chiang, Phelps, & Daw, 2013), group-level behavior reflected a mixture of both strategies. A logistic regression revealed a significant main effect of reward ($p<.0001$), indicating model-free learning, and an interaction between reward and transition type ($p<.01$), the signature of model-based contributions (the third through fifth coefficients in Table 2).

To visualize the relationship between susceptibility to Stroop interference and model-based contribution to behavior, we computed a Model-Based Index for each subject (the individual's coefficient estimate for the previous reward × transition type interaction as in Figure 1B), and plotted this index as a function of infrequent IE. Figure 2 suggests that an individual's susceptibility to Stroop interference (i.e., more slowing, interpreted as poorer cognitive control) negatively predicts the contribution of model-based RL. Statistically, in testing infrequent IE as a linear covariate modulating model-based choice, we found a significant three-way interaction between infrequent IE, reward, and transition type ($p<.05$, Table 2), revealing that Stroop interference predicts a decreased model-based choice contribution. There was no significant effect of infrequent IE upon previous reward ($p=.307$)

suggesting that the predictive effect of susceptibility to response conflict was limited to model-based rather than model-free choice contribution.

The finding that individual susceptibility to Stroop interference negatively predicts the contribution of model-based learning to choices suggests an underlying correspondence between cognitive control ability and the relative expression of dual-systems RL. But a less specific account of performance could in principle explain the cross-task correlation—namely, some more generalized performance variation, such as in gross motivation or task attentiveness, might manifest in a pattern where generally poor-performing subjects showed both larger IEs and less reliance on the more cognitively demanding (model-based) strategy on the RL task. If this were the case, we would expect that IE in both Stroop conditions (infrequent and frequent) would similarly predict model-based choice. However, we found no significant (or even negatively trending) relationship between frequent IE and model-based choice ($p$=.143, Figure 2B) and this relationship was significantly different from the effect of infrequent IE, ($p$<.05, linear contrast), demonstrating a specificity in the across-task relationship and mitigating against an account in terms of generalized performance variation.

Experiment 2 examines individual differences in context processing more precisely, revealing how utilization of task-relevant contextual information—a more specific hallmark of cognitive control (Braver, Barch, & Cohen, 1999)—predicts model-based tendencies in choice behavior. This provides a complimentary demonstration of the relationship between cognitive control and RL, using a structurally different task termed the AX-CPT (Cohen et al., 1999), which has been used to understand context processing deficits in a variety of populations (Braver et al., 2005; Servan-Schreiber, Cohen, & Steingard, 1996). In doing so, we also probe the more general task engagement account because the AX-CPT includes a condition in which successful cognitive control may actually hinder performance.

## Experiment 2

In Experiment 2 we examined the relationship between model-based choice and cognitive control more directly. Fifty-one participants completed the Dot Pattern Expectancy task (MacDonald, 2008), which is structurally equivalent to the AX-Continuous Performance Task (AX-CPT) (Cohen et al., 1999). In each trial, participants were briefly presented with a cue (a blue dot pattern, A or B), followed by a delay period and then a probe (a white dot pattern, X or Y) (Figure 3). Participants are required to make a “target” response only when they see a valid cue-probe pair, referred to as an AX pair. For all other cue-probe sequences participants are instructed to make a “non-target” response. The invalid, non-target cues and probes are called B and Y, respectively, yielding four trial types of interest: AX, AY, BX, and BY. Because the AX trial occurs by far most frequently (more than 2/3 of trials), it engenders a preparatory context triggered by the A cue, and a strong target response bias triggered by the X probe. These effects manifest in difficulty on AY and BX trials, relative to AX (the prepotent target) and BY (where neither type of bias is present).

In the DPX task, the BX trials index the beneficial effects of context on behavior (provided by the cue stimulus), because the X stimulus, considered in isolation, evokes inappropriate, stimulus-driven “target” responding. Utilizing the B context supports good performance

(fast and accurate non-target responses) here, because the cue-driven expectancy can be used to inhibit the incorrect target response (Braver & Cohen, 2000). Importantly, contextual preparation has opposite effects on BX and AY trials, since the predominant X cue, requiring a target response, is even more likely following A. For this reason although the usage of context improves performance on BX trials, it tends to impair performance on AY trials; conversely, less utilization of contextual information (or more probe-driven behavior) causes BX performance to suffer but results in improved AY performance. Thus, performance measures on the two trials, in effect, provide mutual controls for each other. Accordingly, we sought to examine how RTs on correct AY and BX trials, which both index utilization of contextual information (Braver et al., 2005), predict model-based contributions in the Two-stage RL task.

DPX accuracy and RTs at the group level (Table 3) mirrored those of healthy participants in previous work (Henderson et al., 2012)—that is, subjects made faster and more accurate responses on target AX trials and non-target BY trials compared to AY and BX trials (Cue X Probe Accuracy Interaction on $F(1,49)=78.54$, $p<.001$, RT Interaction $F(1,39)=32.65$, $p<.01$). Overall, $d'$-*context*, which provides an estimate of sensitivity corrected for response bias, mirrored that of controls in previous studies using the DPX task (Henderson et al., 2012), [$M=3.32$, $SD=0.76$]. Finally, AY and BX RTs manifested a moderate, negative correlation ($r(49) = -0.41$, $p<.01$) in line with the antagonistic relationship between the two performance measures.

In the two-stage RL task, group-level behavior revealed the same mixture of model-free and model-based strategies (Table 4) as Experiment 1. We first examined if BX RTs—which, in the DPX, should be smaller for individuals using contextual information—predicts model-based contributions to choice. Plotting the Model-based Index as a function of BX RT (Figure 4A) suggests that individuals who were worse performers (slower relative RTs) on BX trials exhibited diminished model-based choice contributions. Put another way, successful BX performance requires inhibition of the prepotent response, the same sort of inhibition required in incongruent Stroop trials in Experiment 1 (Cohen et al., 1999). Interpreted this way, the BX relationship corroborates the Stroop effect relationship reported in Experiment 1 (Figure 2A). Indeed, the full trial-by-trial multilevel logistic regression (Table 4) indicates that BX RT negatively and significantly predicted model-based choice (i.e., the interaction between BX RT and the reward × transition interaction, the signature of model-based learning, was significant).

We hypothesized that AY performance should also predict model-based choice, but that the effect should have the opposite (positive) direction because on these trials, increased use of the contextual information interferes with performance, producing larger RTs. Note that under the hypothesis that all our effects are driven by some more generalized, confounding performance variation, such as differences in gross motivation or engagement between participants, predicts the opposite effect: poorer performance on the AY trials, like Stroop and BX, should track *decreasing* model-based contributions. This relationship is plotted in Figure 4B. We indeed found a significant positive relationship between AY RT and the reward × transition interaction term (the full model coefficients are reported in Table 5). As in Experiment 1, neither AY nor BX RTs interacted significantly with previous reward

(Tables 4 and 5), suggesting that the locus of the predictive effect of the cognitive control variables was in the model-based, rather than the model-free, domain.

Finally, we considered the effects of both of these measures of control (AY and BX RTs) together in a multiple regression. To the extent that these effects may both reflect a common underlying control process or tradeoff, we might expect that there would not be enough unique variance to attribute the relationship with model-based reinforcement learning uniquely to one or the other. Accordingly, their effects upon model-based signatures (expressed as the three-way interactions of AY and BX RT with reward X transition) did not reach significance individually in the presence of one another (Table 6). However, the contrast between these effects (equivalently, in a linear model, the interaction of reward X transition with the average of the two cognitive control measures, AY and BX RTs, or with the differential score AY-BX, similar to Braver et al.'s (2009) proactive RT index) was significant ($p$=.015 Figure 4C). Mirroring the separate regressions above, the contrast between AY and BX RT's interactions with reward was not significant ($p > .5$), suggesting that these predictive effects were limited to the model-based domain. We also attempted the foregoing analyses using BX and AY error rates instead of RTs as predictor variables, finding no significant relationships upon choice behavior, but we suspect this was due to insufficient variance and/or floor effects in these measures.

## Discussion

We probed the connection between cognitive control and the behavioral expression of model-based RL in sequential choice. We examined individual differences in measures of cognitive control across separate task domains to reveal a correspondence between behavioral signatures of the two well-established constructs. In short, we found that individuals exhibiting more goal-related usage of contextual cues—a hallmark of cognitive control (Braver & Barch, 2002)—express a larger contribution of deliberative, model-based strategies, which dominate over simpler habitual (or model-free) behavior in sequential choice. Put another way, difficulty in producing appropriate actions in the face of prepotent, interfering tendencies, operationalized separately across two different experiments via response latencies, was associated with diminished model-based choice signatures in a separate decision-making task. The two experiments reveal a previously unexamined, but intuitive correspondence between two qualitatively different behavioral repertoires—cognitive control and model-based RL.

### Individual Differences in Control and RL Strategies

It is worth noting that the relationship between cognitive control and choice observed here cannot be explained merely in terms of global task engagement, whereby more attentive or grossly motivated individuals bring their full resources to bear on both tasks. This is because utilizing contextual information in the DPX is not globally beneficial, but instead, results in better performance in some circumstances (BX trials) but worse performance in other circumstances (AY trials). Accordingly, subjects who exhibited the slowest RTs in AY trials actually exhibited the strongest contributions of model-based choice behavior in the two-step RL task (Figure 4A). In other words, contrary what is expected from nonspecific

performance differences, poorer behavior in one type of situation in the DPX task predicted better behavior in the sequential choice task (whereas, of course, the complementary pattern of performance was also observed, for BX and Stroop interference). A panoply of recent studies demonstrate how manipulations (e.g., working-memory load, acute stress, disruption of neural activity; Otto et al., 2013a, 2013b; Smittenaar et al., 2013) or individual differences such as age (Eppinger, Walter, Heekeren, & Li, 2013), psychiatric disorders (Voon et al. in press), or personality traits (Skatova, Chan, & Daw, 2013) all attenuate model-based behavioral contributions. The present result highlights the specificity of this choice measure, as being distinct from performance more generally.

Relatedly, individual differences in both task domains is arguably best understood not as reflecting better or worse (or overall more or less motivated) performance per se, but instead as reflecting different strategies or modes of performing. Indeed, model-free RL is a separate learning strategy from model-based RL, which can be overall beneficial (Daw et al., 2005). Similarly, the Dual Mechanisms of Control theory (Braver, 2012) fractionates cognitive control—like the dual-systems RL framework—into two distinct operating modes: proactive control is conceptualized as the sustained maintenance of context information in order to bias attention and action in a goal-driven manner, whereas reactive control is conceived as stimulus-driven, transient control recruited as needed.

On this view, larger interference costs on the Stroop task (when incongruent stimuli are uncommon) in Experiment 1 may reflect reliance upon reactive (rather than a proactive) control to support correct performance on incongruent trials (Grandjean et al., 2012). Moreover, the AX-CPT task (or the DPX as used here in Experiment 2) has been interpreted as more directly dissociating the contributions of reactive and proactive control (Braver, 2012). In particular, a proactive control mode would support the effects of representing the A or B context (harming performance on AY trails and facilitating it on BX trials), whereas in the absence of such cue-triggered expectancies (i.e., to the extent subjects rely on purely reactive control), BX trials evoke inappropriate response tendencies that, as in Stroop, would require reactive control to override them, resulting in slowed responses (Paxton, Barch, Racine, & Braver, 2008). Neurally, the fractionation of the two control modes occurs on the basis of temporal dynamics of activity: proactive control is accompanied by sustained, probe-evoked activation of the DLPFC while reactive control is associated with transient activation of the DLFPC in response to the probe (Braver, Paxton, Locke, & Barch, 2009; Paxton et al., 2008).

Together with the present results, the proactive/reactive interpretation of the AX-CPT data suggests that subjects' more specific reliance on *proactive*, rather than reactive, control is what predicts the ability to carry out a model-based strategy in our sequential choice task. Thus, larger AY RTs, indicative of a proactive strategy, predict more usage of model-based RL, whereas larger BX RTs and infrequent-condition Stroop Interference Effects, indicative of the reactive strategy, predict less model-based strategy usage. As proactive control is characterized by its reliance on contextual information about the preceding cues (Braver, 2012), and a similar sort of higher-order representation (here, of the internal model) is required to carry out model-based choice, we believe that the Dual Mechanisms of Control theory provides an intuitive framework for understanding the observed patterns of behavior.

Such an interpretation would refine the relatively broad account that cognitive control, more generally, is needed to support model-based action.

However, at least two questions remain. First, it is not entirely clear on this view why reactive control (which also supports correct responding on AY and incongruent trials, albeit at an RT cost, perhaps by retrieving the goal information while suppressing the prepotent response) would not also be effective in enabling model-based responding. It may be that there is nothing about the RL choice situation – analogous to an incongruent cue – that specifically evokes or engages reactive control, so that a tendency to utilize the internal model must in this setting be proactively generated.

A second concern is that it might have been expected that in the Stroop task, proactive control is engaged most strongly when incongruent stimuli are frequent, (Bugg et al., 2011). Thus we might have expected that interference effects in the incongruent-frequent condition —diminished by the engagement of proactive control—would, in turn, negatively correlate with model-based strategy usage in the RL task, potentially even more strongly than in the incongruent-frequent condition. One possibility is that under the conditions of our study, good frequent-incongruent performance was more driven by stimulus-response learning about the individual incongruent items rather than global, proactive control (Jacoby et al. 2003; Bugg et al. 2008). Since such learning is itself similar to model-free RL, which presumably competes with model-based RL, this interpretation might be consistent with the trend toward the opposite effect (faster incongruent-frequent RTs, interpreted as better stimulus-response learning, tracking worse model-based RL) in our data. However, a full test of this interpretation would require separately manipulating list-level and item-level incongruency (Jacoby et al. 2003; Bugg et al. 2008) to dissociate global strategic adjustments in control from associative learning.

## Neural Substrates of Cognitive Control and RL

Interestingly, human neuroimaging and primate neurophysiology work suggest that the brain regions critical in cognitive control and goal-directed choice span nearby sections of the medial wall of the PFC. In particular, BOLD activity in the anterior cingulate cortex (ACC) is well-documented to accompany error or conflict in tasks like the Stroop (Kerns et al., 2004) and has been argued to be implicated in allocation of control (e.g., via conflict or error monitoring, Botvinick, Cohen, & Carter, 2004). Strikingly, Alexander and Brown (2011) propose a computational model of these responses in which they reflect action-outcome learning of a sort essentially equivalent to model-based RL. Meanwhile, value-related activity in RL and decision tasks is widely reported in a nearby strip of PFC extending from adjacent rostral cingulate and medial PFC down through ventromedial and medial orbitofrontal PFC (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Hare, Camerer, & Rangel, 2009). This activity has been suggested to underlie goal-directed or model-based learning, because it is sensitive to devaluation and contingency manipulation (Valentin, Dickinson, & O'Doherty, 2007) and more abstract model-based inference (Daw et al., 2011; Hampton, Bossaerts, & O'Doherty, 2006). Meanwhile, in rodents, lesions in prelimbic cortex (a potentially homologous ventromedial prefrontal area) abolish goal-directed behavior (Balleine & O'Doherty, 2009), and in primates, neurons in the orbitofrontal cortex

encode goal-specific utilities, divorced from sensory or motor representations (Padoa-Schioppa & Assad, 2006).

Recent neuroscientific work also hints at a more lateral neural substrate underpinning both cognitive control and model-based RL. Studies utilizing multi-step choice tasks highlight a crucial role for the dorsolateral (DLPFC) and lateral prefrontal cortex in decision-makers' learning and utilization of the model of the environment—the defining characteristic of model-based RL (Gläscher et al., 2010; Lee, Shimojo, & O'Doherty, 2014). Further, Smittenaar et al. (2013) demonstrated that disruption of DLFPC activity selectively attenuates model-based choice. Similarly, in simple context-processing tasks, DLPFC activation accompanies utilization of goal-related cues—the hallmark of cognitive control (Braver, Paxton, Locke, & Barch, 2009; Egner & Hirsch, 2005; Kerns et al., 2004), while disruption of activity in the same region can impair context-dependent responding (D'Ardenne et al., 2012).

## Relationships Between the Computational Mechanisms of Control And RL

Both RL and cognitive control have been the subject of intensive computational modeling, and to the extent that these two constructs overlap, these theories may be relevant across both domains. Theoretical accounts of dual-systems RL tend to conceptualize the arbitration question—that is, how the brain decides, at choice time, whether to make the choice preferred by the model-based versus model-free system—as an even-handed competition between the two systems. Abstractly, this choice has been conceptualized as involving a cost-benefit trade-off in their flexibility and computational expense (Daw, Niv, & Dayan, 2005; Keramati, Dezfouli, & Piray, 2011), an idea which explains considerable data about the task circumstances under which rodents tend to exhibit either strategy. Little explanation, however, has been offered about this competition at a process level. Considering the arbitration question in light of the cognitive control framework—and moreover, the present results—suggests that rather than selection between two preferences, the interaction between the two systems may be more top-down or hierarchical in nature. In such an arbitration scheme, cognitive control might bolster the influence of model-based relative to model-free choice by actively boosting the influence of task-relevant representations or actively inhibiting model-free responses. On this view, responses favored by the model-free system (such as repeating a previously unrewarded response following a rare transition, Figure 1C) are akin to prepotent (inappropriate) color-reading responses in the Stroop task or "target" responses to BX stimuli in the DPX task. Meanwhile, the choices that would be best taking into account the internal model of the task structure (akin to contextual or goal-related information in control tasks) must override them. Indeed, recent functional neuroimaging work finds support for such an inhibition scheme (Lee, Shimojo, & O'Doherty, 2014).

Computationally, this view invites reconsideration of the RL arbitration problem in terms of the mechanisms modeled in the cognitive control literature, by which higher-order representations are strengthened and/or responses are delayed or inhibited to favor controlled responding (Cohen, Barch, Carter, & Servan-Schreiber, 1999; Shenhav, Botvinick, & Cohen, 2013). This points to an altogether different and more detailed process-

level account of competition than suggested on previous RL work, which might speak particularly to the within-trial time dynamics of the tradeoff during the choice process (see also Solway & Botvinick, 2012). Importantly, whereas model-free action values are directly produced by learning and can simply be retrieved at choice time (consistent with prepotency), model-based values are typically viewed as computed at decision time, through a sort of mental simulation using the internal model. Since the latter process takes time, it is naturally matched to control theories envisioning a race to suppress a prepotent response while a more controlled process builds up (Frank, 2006).

## Possible Future Directions

Another implication of a cognitive control perspective for RL is that the cognitive control literature has also focused extensively on trial-to-trial shifts in the engagement of control, and the related phenomenology of sequential effects such as post-error slowing (Botvinick, Braver, Barch, Carter, & Cohen, 2001). In the same fashion, it is possible that the tendency toward model-based control shifts from trial to trial as a function of observed stability in rewards or transition structure as it they may signal a need for increased control of a proactive sort. Such trial-trial adjustments have received relatively little attention in RL so far, but hints of adjustments in reliance on model-based RL have been observed in sequential choice behavior (Lee et al., 2014; Simon & Daw, 2011). Future work, following the tradition of the cognitive control literature, should aim to uncover more precisely 1) what sorts of events in the environment trigger these shifts and 2) how these shifts may be implemented neurally.

Conversely, and again suggesting future theoretical and experimental work, decision-theoretic treatments of how the brain balances the costs and benefits of model-based control in choosing an RL strategy (Daw et al., 2005; Keramati et al., 2011; Pezzulo, Rigoli, & Chersi, 2013) can potentially be applied to understanding the costs and benefits of cognitive control more generally. In particular, this approach might serve as the basis for an analogous account of how the brain chooses a more proactive or reactive control strategy under different circumstances. Relatedly, recent work demonstrates that people treat as costly the subjective “effort” of controlled behavior such as task switching or working memory demand (Kool et al 2010; Westbrook, Kester, & Braver, 2013). Individual differences in the disinclination toward such mental effort might be key source driving the corresponding tendencies to use both model-based RL and proactive control across our tasks. Testing this idea, of course, would require additionally assessing individual differences in subjective effort cost in a study like the current one.

Separate from the issues of competition considered thus far, another complementary relationship between theories RL and cognitive control concerns the role of learning in establishing the sorts of higher-level contextual representations or task sets that are supposed to be privileged by control. Recent research (Collins & Frank, 2013; Collins & Koechlin, 2012; Gershman & Niv 2010) aims to extend principles of RL, hierarchically, to learning at this level.

### Broader Implications

Finally, characterizing and understanding the correspondence between cognitive control and decision-making may be of practical importance. Prominent accounts of substance abuse ascribe compulsive and drug-seeking behaviors to the aberrant expression of habitual or stimulus-driven control systems at the expense of goal-directed action (Everitt & Robbins, 2005), respectively instantiated in dual-systems RL as the model-free and model-based systems. At the same time, impairments in such response inhibition are observed in populations showing pathologically compulsive choices such as cocaine abusers and pathological gamblers (Odlaug, Chamberlain, Kim, Schreiber, & Grant, 2011; Volkow et al., 2010). Within the dual modes of control framework, these impairments are thought to stem from the breakdown of the proactive, rather than reactive control system (Garavan, 2011). The finding that individuals—within a non-clinical population—who exhibit more difficulty inhibiting stimulus-driven responding also show more reliance upon predominately model-free choice dovetails neatly with both of these accounts.

## Acknowledgments

We thank John McDonnell for assistance with online data collection and Deanna Barch and Jonathan Cohen for helpful conversations.

## References

Alexander WH, Brown JW. Medial prefrontal cortex as an action-outcome predictor. Nature Neuroscience. 2011; 14(10):1338–1344.10.1038/nn.2921

Balleine B, O'Doherty J. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. Neuropsychopharmacology. 2009; 35(1):48–69.10.1038/npp.2009.131 [PubMed: 19776734]

Besner D, Stolz JA, Boutilier C. The stroop effect and the myth of automaticity. Psychonomic Bulletin & Review. 1997; 4(2):221–225.10.3758/BF03209396 [PubMed: 21331828]

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. Conflict monitoring and cognitive control. Psychological Review. 2001; 108(3):624–652.10.1037/0033-295X.108.3.624 [PubMed: 11488380]

Botvinick MM, Cohen JD, Carter CS. Conflict monitoring and anterior cingulate cortex: an update. Trends in Cognitive Sciences. 2004; 8(12):539–546.10.1016/j.tics.2004.10.003 [PubMed: 15556023]

Braver TS. The variable nature of cognitive control: a dual mechanisms framework. Trends in Cognitive Sciences. 2012; 16(2):106–113.10.1016/j.tics.2011.12.010 [PubMed: 22245618]

Braver TS, Barch DM. A theory of cognitive control, aging cognition, and neuromodulation. Neuroscience & Biobehavioral Reviews. 2002; 26(7):809–817.10.1016/S0149-7634(02)00067-2 [PubMed: 12470692]

Braver TS, Barch DM, Cohen JD. Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. Biological Psychiatry. 1999; 46(3):312–328.10.1016/S0006-3223(99)00116-X [PubMed: 10435197]

Braver TS, Cohen JD. On the control of control: The role of dopamine in regulating prefrontal function and working memory. Control of Cognitive Processes: Attention and Performance. 2000; XVIII:713–737.

Braver TS, Paxton JL, Locke HS, Barch DM. Flexible neural mechanisms of cognitive control within human prefrontal cortex. Proceedings of the National Academy of Sciences. 2009; 106(18):7351–7356.10.1073/pnas.0808187106

Braver TS, Satpute AB, Rush BK, Racine CA, Barch DM. Context Processing and Context Maintenance in Healthy Aging and Early Stage Dementia of the Alzheimer's Type. Psychology and Aging. 2005; 20(1):33–46.10.1037/0882-7974.20.1.33 [PubMed: 15769212]

Bugg JM, McDaniel MA, Scullin MK, Braver TS. Revealing list-level control in the Stroop task by uncovering its benefits and a cost. Journal of Experimental Psychology: Human Perception and Performance. 2011; 37(5):1595–1606.10.1037/a0024670 [PubMed: 21767049]

Carter CS, Macdonald AM, Botvinick M, Ross LL, Stenger VA, Noll D, Cohen JD. Parsing executive processes: Strategic vs. evaluative functions of the anterior cingulate cortex. PNAS. 2000; 97(4): 1944–1948.10.1073/pnas.97.4.1944 [PubMed: 10677559]

Chatham CH, Frank MJ, Munakata Y. Pupillometric and behavioral markers of a developmental shift in the temporal dynamics of cognitive control. Proceedings of the National Academy of Sciences. 2009; 106(14):5529–5533.10.1073/pnas.0810002106

Cohen JD, Barch DM, Carter C, Servan-Schreiber D. Context-processing deficits in schizophrenia: Converging evidence from three theoretically motivated cognitive tasks. Journal of Abnormal Psychology. 1999; 108(1):120–133.10.1037/0021-843X.108.1.120 [PubMed: 10066998]

Collins AGE, Frank MJ. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychological Review. 2013; 120(1):190–229.10.1037/a0030852 [PubMed: 23356780]

Crump MJC, McDonnell JV, Gureckis TM. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PLoS ONE. 2013; 8(3):e57410.10.1371/journal.pone.0057410 [PubMed: 23516406]

D'Ardenne K, Eshel N, Luka J, Lenartowicz A, Nystrom LE, Cohen JD. Role of prefrontal cortex and the midbrain dopamine system in working memory updating. PNAS. 2012; 109(49):19900–19909.10.1073/pnas.1116727109 [PubMed: 23086162]

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron. 2011; 69(6):1204–1215.10.1016/j.neuron.2011.02.027 [PubMed: 21435563]

Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8(12):1704–1711.10.1038/nn1560 [PubMed: 16286932]

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441:876–879. [PubMed: 16778890]

Dolan RJ, Dayan P. Goals and Habits in the Brain. Neuron. 2013; 80(2):312–325.10.1016/j.neuron.2013.09.007 [PubMed: 24139036]

Economides M, Guitart-Masip M, Kurth-Nelson Z, Dolan RJ. Anterior Cingulate Cortex Instigates Adaptive Switches in Choice by Integrating Immediate and Delayed Components of Value in Ventromedial Prefrontal Cortex. The Journal of Neuroscience. 2014; 34(9):3340–3349.10.1523/JNEUROSCI.4313-13.2014 [PubMed: 24573291]

Egner T, Hirsch J. The neural correlates and functional integration of cognitive control in a Stroop task. NeuroImage. 2005; 24(2):539–547.10.1016/j.neuroimage.2004.09.007 [PubMed: 15627596]

Eppinger B, Walter M, Heekeren HR, Li S-C. Of goals and habits: Age-related and individual differences in goal-directed decision-making. Frontiers in Neuroscience. 2013; 7(253)10.3389/fnins.2013.00253

Everitt BJ, Robbins TW. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nature Neuroscience. 2005; 8:1481–1489.10.1038/nn1579

Frank MJ. Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. Neural Networks. 2006; 19(8):1120–1136.10.1016/j.neunet.2006.03.006 [PubMed: 16945502]

Garavan, H. Impulsivity and Addiction. In: Adinoff, B.; Stein, EA., editors. Neuroimaging in Addiction. John Wiley & Sons, Ltd; 2011. p. 157-176.

Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron. 2010; 66(4):585–595.10.1016/j.neuron.2010.04.016 [PubMed: 20510862]

Grandjean J, D'Ostilio K, Phillips C, Balteau E, Degueldre C, Luxen A, Collette F. Modulation of Brain Activity during a Stroop Inhibitory Task by the Kind of Cognitive Control Required. PLoS ONE. 2012; 7(7):e41513.10.1371/journal.pone.0041513 [PubMed: 22911806]

Hampton AN, Bossaerts P, O'Doherty JP. The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. J Neurosci. 2006; 26(32):8360–8367.10.1523/JNEUROSCI.1010-06.2006 [PubMed: 16899731]

Hare TA, Camerer CF, Rangel A. Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. Science. 2009; 324(5927):646–648.10.1126/science.1168450 [PubMed: 19407204]

Henderson D, Poppe AB, Barch DM, Carter CS, Gold JM, Ragland JD, MacDonald AW. Optimization of a Goal Maintenance Task for Use in Clinical Applications. Schizophrenia Bulletin. 2012; 38(1): 104–113.10.1093/schbul/sbr172 [PubMed: 22199092]

Højsgaard, S.; Halekoh, U. doBy: Groupwise computations of summary statistics, general linear contrasts and other utilities. 2009. Retrieved from http://CRAN.R-project.org/package=doBy

Holroyd CB, Yeung N. Motivation of extended behaviors by anterior cingulate cortex. Trends in Cognitive Sciences. 2012; 16(2):122–128.10.1016/j.tics.2011.12.008 [PubMed: 22226543]

Kahneman, D. Thinking, Fast and Slow. Macmillan; 2011.

Kane MJ, Engle RW. Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. Journal of Experimental Psychology: General. 2003; 132(1):47–70.10.1037/0096-3445.132.1.47 [PubMed: 12656297]

Keramati M, Dezfouli A, Piray P. Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. PLoS Comput Biol. 2011; 7(5):e1002055.10.1371/journal.pcbi.1002055 [PubMed: 21637741]

Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS. Anterior Cingulate Conflict Monitoring and Adjustments in Control. Science. 2004; 303(5660):1023–1026.10.1126/science.1089910 [PubMed: 14963333]

Kool W, McGuire JT, Rosen ZB, Botvinick MM. Decision making and the avoidance of cognitive demand. Journal of Experimental Psychology: General. 2010; 139(4):665–682.10.1037/a0020198 [PubMed: 20853993]

Lee SW, Shimojo S, O'Doherty JP. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. Neuron. 2014; 81(3):687–699.10.1016/j.neuron.2013.11.028 [PubMed: 24507199]

Lindsay DS, Jacoby LL. Stroop process dissociations: The relationship between facilitation and interference. Journal of Experimental Psychology: Human Perception and Performance. 1994; 20(2):219–234.10.1037/0096-1523.20.2.219 [PubMed: 8189189]

Loewenstein G. Out of Control: Visceral Influences on Behavior. Organizational Behavior and Human Decision Processes. 1996; 65(3):272–292.10.1006/obhd.1996.0028

Logan GD, Zbrodoff NJ. When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. Memory & Cognition. 1979; 7(3):166–174.10.3758/BF03197535

MacDonald AW. Building a Clinically Relevant Cognitive Task: Case Study of the AX Paradigm. Schizophrenia Bulletin. 2008; 34(4):619–628.10.1093/schbul/sbn038 [PubMed: 18487225]

Maddox WT, Pacheco J, Reeves M, Zhu B, Schnyer DM. Rule-based and information-integration category learning in normal aging. Neuropsychologia. 2010; 48(10):2998–3008.10.1016/j.neuropsychologia.2010.06.008 [PubMed: 20547171]

McDonnell, JV.; Martin, JB.; Markant, DB.; Coenen, A.; Rich, AS.; Gureckis, TM. psiTurk (Version 1.02) [Software]. New York, NY: New York University; 2012. Retrieved from https://github.com/NYUCCL/psiTurk

Odlaug BL, Chamberlain SR, Kim SW, Schreiber LRN, Grant JE. A neurocognitive comparison of cognitive flexibility and response inhibition in gamblers with varying degrees of clinical severity. Psychological Medicine. 2011; 41(10):2111–2119.10.1017/S0033291711000316 [PubMed: 21426627]

Otto AR, Gershman SJ, Markman AB, Daw ND. The Curse of Planning Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. Psychological Science. 2013; 24(5):751–761.10.1177/0956797612463080 [PubMed: 23558545]

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working-memory capacity protects model-based learning from stress. Proceedings of the National Academy of Sciences. 2013; 110(52):20941–20946.10.1073/pnas.1312011110

Padoa-Schioppa C, Assad JA. Neurons in the orbitofrontal cortex encode economic value. Nature. 2006; 441(7090):223–226.10.1038/nature04676 [PubMed: 16633341]

Paxton JL, Barch DM, Racine CA, Braver TS. Cognitive Control, Goal Maintenance, and Prefrontal Function in Healthy Aging. Cerebral Cortex. 2008; 18(5):1010–1028.10.1093/cercor/bhm135 [PubMed: 17804479]

Pezzulo G, Rigoli F, Chersi F. The Mixed Instrumental Controller: Using Value of Information to Combine Habitual Choice and Mental Simulation. Frontiers in Psychology. 2013; 410.3389/fpsyg.2013.00092

Pinheiro, JC.; Bates, DM. Mixed-Effects Models in S and S-PLUS. New York: Springer; 2000.

Ratcliff R. Methods for dealing with reaction time outliers. Psychological Bulletin. 1993; 114(3):510–532.10.1037/0033-2909.114.3.510 [PubMed: 8272468]

Redick TS, Engle RW. Integrating working memory capacity and context-processing views of cognitive control. The Quarterly Journal of Experimental Psychology. 2011; 64(6):1048–1055.10.1080/17470218.2011.577226 [PubMed: 21644190]

Rushworth MFS, Noonan MP, Boorman ED, Walton ME, Behrens TE. Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron. 2011; 70(6):1054–1069.10.1016/j.neuron.2011.05.014 [PubMed: 21689594]

Schultz W, Dayan P, Montague PR. A Neural Substrate of Prediction and Reward. Science. 1997; 275(5306):1593–1599.10.1126/science.275.5306.1593 [PubMed: 9054347]

Servan-Schreiber D, Cohen JD, Steingard S. Schizophrenic deficits in the processing of context: A test of a theoretical model. Archives of General Psychiatry. 1996; 53(12):1105–1112.10.1001/archpsyc.1996.01830120037008 [PubMed: 8956676]

Shenhav A, Botvinick MM, Cohen JD. The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. Neuron. 2013; 79(2):217–240.10.1016/j.neuron.2013.07.007 [PubMed: 23889930]

Simon, DA.; Daw, ND. Environmental statistics and the trade-off between model-based and TD learning in humans. In: Shawe-Taylor, J.; Zemel, RS.; Bartlett, PL.; Pereira, F.; Weinberger, KQ., editors. Advances in Neural Information Processing Systems 24. 2011. p. 127-135.

Skatova A, Chan PA, Daw ND. Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. Frontiers in Human Neuroscience. 2013; 7:525.10.3389/fnhum.2013.00525 [PubMed: 24027514]

Smittenaar P, FitzGerald THB, Romei V, Wright ND, Dolan RJ. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. Neuron. 2013; 80(4):914–919.10.1016/j.neuron.2013.08.009 [PubMed: 24206669]

Valentin VV, Dickinson A, O'Doherty JP. Determining the neural substrates of goal-directed learning in the human brain. The Journal of Neuroscience. 2007; 27(15):4019–4026.10.1523/JNEUROSCI.0564-07.2007 [PubMed: 17428979]

Volkow ND, Fowler JS, Wang GJ, Telang F, Logan J, Jayne M, Swanson JM. Cognitive control of drug craving inhibits brain reward regions in cocaine abusers. NeuroImage. 2010; 49(3):2536–2543.10.1016/j.neuroimage.2009.10.088 [PubMed: 19913102]

Voon V, Derbyshire K, Ruck C, Irvine M, Worbe Y, Enander J, Schreiber L. Disorders of compulsivity: a common bias towards learning habits. Molecular Psychiatry. in press.

Westbrook A, Kester D, Braver TS. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. PloS One. 2013; 8(7):e68210.10.1371/journal.pone.0068210 [PubMed: 23894295]

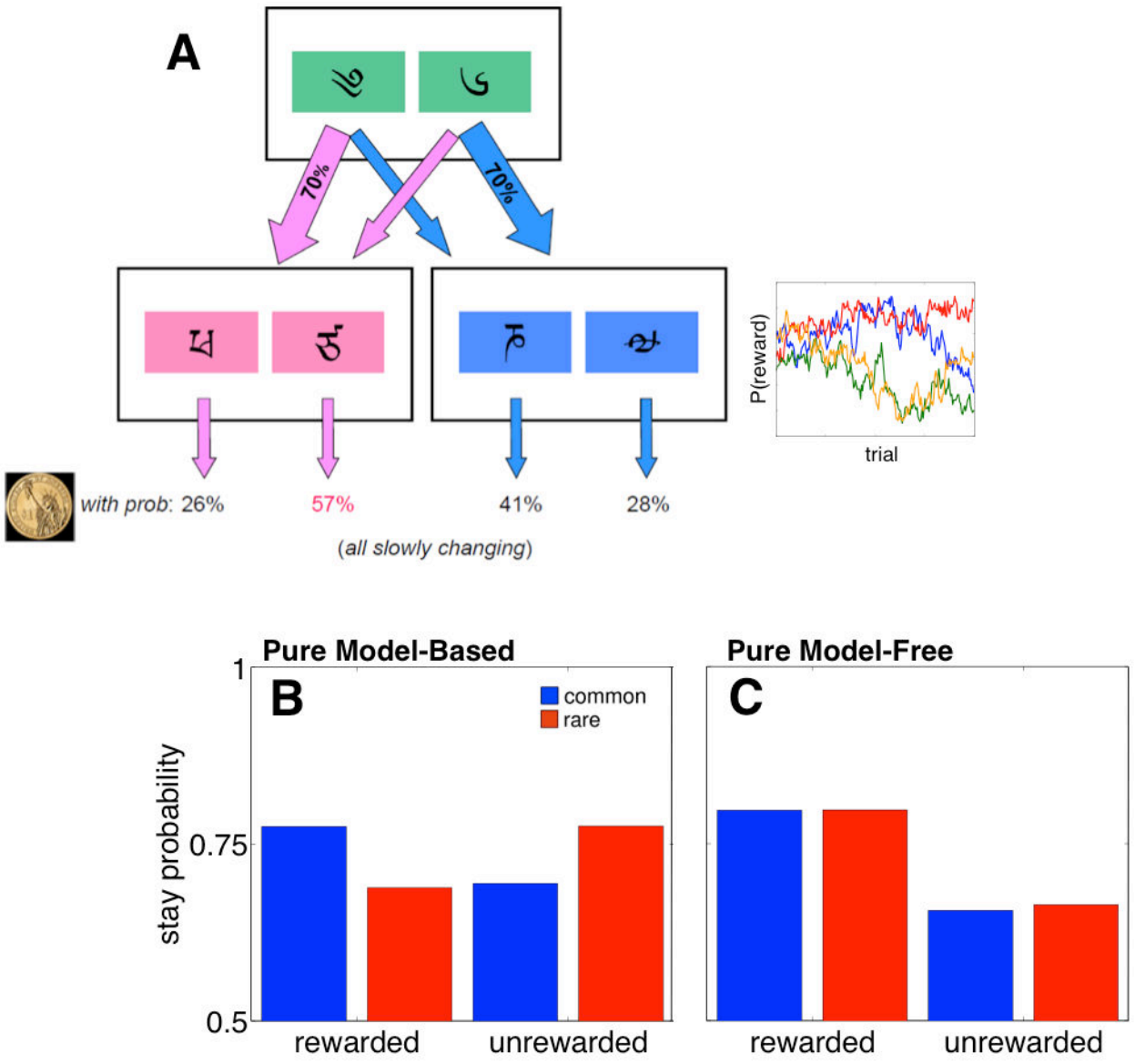


**Figure 1.**

(A) The structure of the Two-Stage RL task. In each trial, subjects chose between two initial options, leading to either of two second-stage choices (pink or blue states), these for different, slowly changing, chances of monetary reward. Each first-stage option led more frequently to one of the second-stage states (a “common” transition), however, on 30% of trials (“rare”) it instead led to the other state. (B) A Model-based choice strategy predicts that rewards after rare transitions should affect the value of the unchosen first-stage option, leading to a predicted interaction between the factors of reward and transition probability. (C) In contrast, a Model-Free strategy predicts that a first-stage choice resulting in reward is more likely to be repeated on the subsequent trial regardless of whether that reward occurred after a common or rare transition.

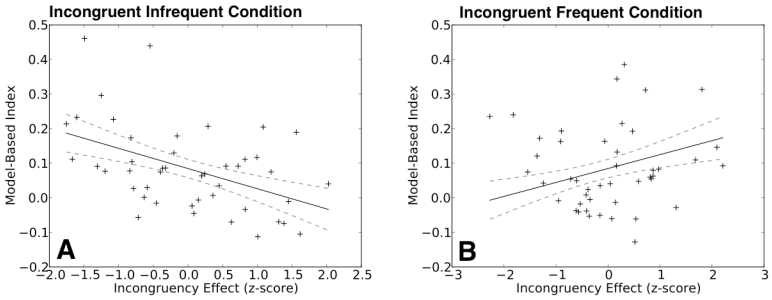


**Figure 2.**

Visualization of relationship between Stroop IE and model-based contributions to choice in Experiment 1. The Model-based index is calculated as individual subjects' model-based effect sizes (arbitrary units) conditional on the group-level mixed-effects logistic regression. (A) Infrequent IE negatively predicts model-based contribution to choice in the Two-stage RL task. (B). Frequent IE effect does not significantly predict model-based choice contribution. Regression lines are computed from the group-level effect of infrequent IE (A) and frequent IE (B). Dashed gray lines indicate standard errors about the regression line, estimated from the group-level mixed effects regression.

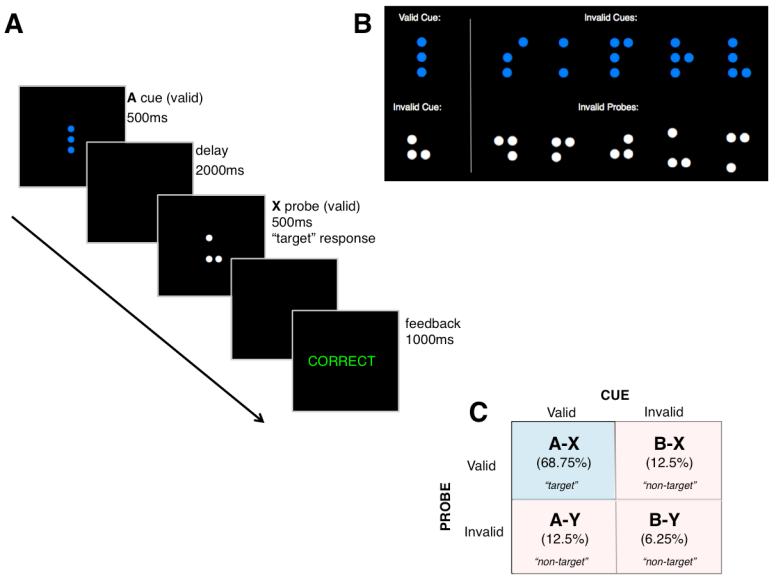


**Figure 3.**

Dot Pattern Expectancy (DPX) Task. (A) An example sequence of cue-probe stimuli and the type of response (target or non-target) required. The valid cue is referred to as "A" and the valid probe is referred to as "X". Non-"A" cues are referred to as "B" cues, and non-"X" probes are referred to as "Y"-probes. Subjects are instructed to make a "target" response only when an "X" probe follows an "A" cue; non-target responses are required for all other stimuli. (B) Stimuli set for DPX task. Cues are always presented in blue while probes are always presented in white. (C) Schema depicting the four trial types, required responses, and presentation rates. See main text for additional task details.

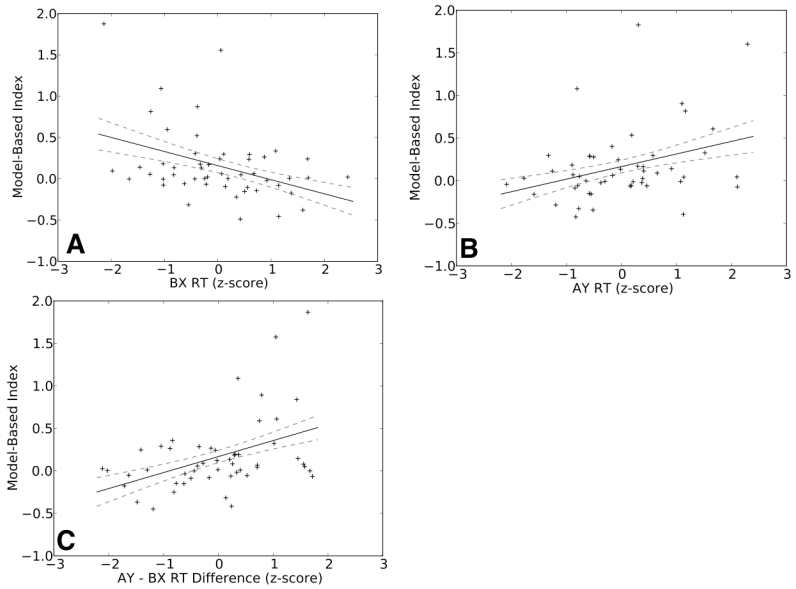


**Figure 4.**

Visualization of relationship between DPX task performance and model-based contributions to choice in Experiment 2. Model-based Index is calculated the same way as in Figure 2. (A) BX RTs negatively predict model-based choice. (B) AY RTs in the DPX task positively predict model-based contributions to choice in the Two-stage RL task. (C) Contrast of AY RT and BX RT effects in the Full model, reveal that unique contribution of AY RTs to prediction of model-based choice contribution. Dashed gray lines indicate standard errors about the regression line, estimated from the group-level mixed effects regression

**Table 1**

Average error Rates and median RTs for Congruent versus Incongruent Trials in the Stroop task.

| Condition | Trial Type | Error Rate (SD) | RT (ms) (SD) |
|---|---|---|---|
| Incongruent Infrequent | Congruent | 0.02 (0.02) | 567.47 (70.86) |
| | Incongruent | 0.08 (0.07) | 697.6 (147.89) |
| Incongruent Frequent | Congruent | 0.04 (0.06) | 573.21 (81.33) |
| | Incongruent | 0.04 (0.04) | 602.07 (88.0) |

## Table 2

Logistic regression coefficients indicating the influence of the Stroop Incongruency Effect (separately for incongruent infrequent and incongruent frequent blocks), outcome of previous trial, and transition type of previous trial, upon response repetition. Asterisks denote significance at the .05 level.

| Coefficient | Estimate (SE) | p-value |
|---|---|---|
| (Intercept) | 0.65 (0.10) | <.0001 |
| stroop (infrequent) | −0.10 (0.11) | 0.361 |
| stroop (frequent) | −0.02 (0.11) | 0.868 |
| reward | 0.16 (0.03) | <.0001 |
| transition | 0.04 (0.02) | 0.116 |
| reward × transition | 0.08 (0.03) | 0.001* |
| reward × stroop (infrequent) | 0.02 (0.03) | 0.648 |
| transition × stroop (infrequent) | −0.02 (0.02) | 0.307 |
| reward × stroop (frequent) | 0.00 (0.03) | 0.933 |
| transition × stroop (frequent) | 0.04 (0.02) | 0.103 |
| reward × transition × stroop (infrequent) | −0.06 (0.03) | 0.031* |
| reward × transition × stroop (frequent) | 0.04 (0.03) | 0.143 |

**Table 3**

Average error Rates and median RTs for the four trial types in the DPX task.

| Trial Type | Error Rate (SD) | RT (ms) (SD) |
|---|---|---|
| AX | 0.02 (0.02) | 543.01 (110.20) |
| AY | 0.13 (0.13) | 686.98 (100.29) |
| BX | 0.15 (0.13) | 482.74 (153.19) |
| BY | 0.02 (0.07) | 507.08 (184.54) |

**Table 4**

Logistic regression coefficients indicating the influence of BX RTs, outcome of previous trial, and transition type of previous trial, upon response repetition.

| Coefficient | Estimate (SE) | p-value |
|---|---|---|
| (Intercept) | 1.57 (0.14) | <2e-16 |
| BX RT | 0.13 (0.14) | 0.324 |
| reward | 0.86 (0.09) | <2e-16 |
| transition | −0.01 (0.04) | 0.791 |
| reward × transition | 0.16 (0.08) | 0.058 |
| BX RT × reward | −0.10 (0.04) | 0.013* |
| BX RT × transition | 0.16 (0.07) | 0.027* |
| BX RT × reward × transition | −0.17 (0.07) | 0.020* |

## Table 5

Logistic regression coefficients indicating the influence of outcome of AY RT, outcome of previous trial, and transition type of previous trial, upon response repetition.

| Coefficient | Estimate (SE) | p-value |
|---|---|---|
| (Intercept) | 1.57 (0.14) | <2e-16 |
| AY RT | 0.09 (0.14) | 0.525 |
| reward | 0.86 (0.09) | <2e-16 |
| transition | −0.01 (0.04) | 0.837 |
| reward × transition | 0.17 (0.07) | 0.027* |
| AY RT × reward | −0.04 (0.09) | 0.672 |
| AY RT × transition | 0.09 (0.04) | 0.018* |
| AY RT × reward × transition | 0.15 (0.07) | 0.045* |

## Table 6

Logistic regression coefficients indicating the influence of AY RT, BX RT, outcome of previous trial, and transition type of previous trial, upon response repetition.

| Coefficient | Estimate (SE) | p-value |
|---|---|---|
| (Intercept) | 1.57 (0.13) | <2e-16 |
| AY RT | 0.86 (0.08) | <2e-16 |
| BX RT | −0.01 (0.04) | 0.864 |
| reward | 0.17 (0.07) | 0.022* |
| transition | 0.18 (0.15) | 0.234 |
| reward × transition | 0.21 (0.15) | 0.157 |
| AY RT × reward | 0.04 (0.09) | 0.679 |
| AY RT × transition | 0.06 (0.04) | 0.142 |
| BX RT × reward | 0.18 (0.09) | 0.058 |
| BX RT × transition | −0.06 (0.04) | 0.135 |
| AY RT × reward × transition | 0.09 (0.08) | 0.244 |
| BX RT × reward × transition | −0.13 (0.08) | 0.115 |