Control adjustment costs limit goal flexibility: Empirical evidence and a theoretical account

Ivan Grahek*[,1], Xiamin Leng*[,1], Sebastian Musslick[1,2], & Amitai Shenhav[1]

[1]*Department of Cognitive, Linguistic, and Psychological Sciences; Carney Institute for Brain Science; Brown University; Providence, RI, USA*

[2]*Institute of Cognitive Science; Osnabrück University; Osnabrück, Germany*

*These authors contributed equally to this work

* Corresponding authors. E-mail address: ivan_grahek@brown.edu & xiamin_leng@brown.edu

Abstract

The ability to switch between goals is the cornerstone of human cognition and behavior. Cognitive control allows for rapid adjustments of cognition in accordance with new goals, but control adjustments come at a cost. This cost is traditionally studied in situations that demand changes to one's task, without necessitating other changes in the control state. Goal flexibility, however, often entails maintaining the same task while adjusting the amount and type of control being allocated to that task. For instance, different stages of a given task might require us to process information more or less efficiently (e.g., by varying levels of attention) and/or respond more or less cautiously (e.g., by varying response thresholds). Across four experiments, we show that such within-task control adjustments incur a performance cost, and that a dynamical systems model can explain the source of these costs. Participants performed a single cognitively demanding task (the color-word Stroop) under varying performance goals (e.g., to be fast or to be accurate). We modeled control allocation to include a dynamic process of adjusting from one's current control state to a target state for a given performance goal. By incorporating inertia into this adjustment process, our model predicts and our empirical findings confirm that people will under-shoot their target control state more (i.e., exhibit larger adjustment costs) when (a) goals switch rather than remain fixed over a block (Study 1); (b) target control states are more distant from one another (Study 2); (c) less time is given to adjust to the new goal (Study 3); and (d) when anticipating having to switch goals more frequently (Study 4). Our findings demonstrate that there is a cost to adjusting control to meet one's goal – even in the absence of a task change – and show that this cost can emerge directly from the dynamics of control adjustment. In so doing, they shed new light on the sources of and constraints on flexibility in human goal-directed behavior.

Keywords: cognitive control; goal flexibility; inertia; motivation; drift diffusion model

Introduction

Every day we pursue hundreds of goals. As we write e-mails, participate in meetings, and schedule weekend trips, we flexibly adjust our approach to processing information (i.e., the configuration of our cognitive control) to meet these goals. For instance, we might choose our words more carefully when responding to an e-mail from a supervisor than when responding to a friend or colleague. The ability to maintain goals and flexibly switch between them is a cornerstone of human cognition (Austin & Vancouver, 1996; Golwitzer, 2012; Dolan & Dayan, 2013; Haggard, 2019; O'Reilly, 2020), and is essential for adaptive behavior in changing environments (Goschke, 2013; Musslick & Cohen, 2021). Yet, switching between different goals appears to come at a cost, manifesting as poorer performance and less progress towards each individual goal relative to what would be achieved in the absence of such switches. Such costs are more pronounced in both psychopathology and frontal lobe damage, resulting in severe impairments to cognitive and behavioral flexibility (Shallice & Burgess, 1991; Kashdan & Rottenberg, 2010; Millan et al., 2012). In spite of how central the cost of switching between goals is to daily function in health and disease, its origin remains poorly understood.

The process of switching task goals has been subject to the earliest investigations of modern cognitive psychology (Jersild, 1927). Central to this process is the concept of cognitive control, which enables us to flexibly adjust information processing in service of the current task goal (Miyake et al., 2000; Diamond, 2013; Botvinick & Cohen, 2014). Control processes underlying the switching of goals are commonly studied in the context of changing task sets, whereby participants must adapt different rules for processing stimuli and responding depending on the current goal (Alport et al., 1994; Rogers & Monsell, 1995). The key finding is that switches between task sets are associated with a cost in performance (e.g., lower error rates

and/or slower reaction times) compared to task repetitions. Such experiments have been extremely fruitful for understanding the costs of switching between discrete stimulus-response contingencies (Kiesel et al., 2010; Koch et al., 2018), such as costs associated with reconfiguring control signals between tasks. However, as the e-mail-writing example makes clear, flexible adjustments between goals can occur in the absence of task switches. Even as we perform the same task – whether writing e-mails, editing manuscripts, or conversing with different people over lunch – we may be induced to adjust how we process and respond to information (e.g., with what level of focus or caution). Whether such changes in performance goals incur the same costs as switches between discrete task-sets remains an open question. To better understand the process and the constraints by which people shift between goals, it is thus necessary to study situations in which goals change, while other aspects of the task are held constant.

Here, we seek to fill this gap by developing a theoretical framework to understand the costs of adjusting cognitive control states, and a behavioral paradigm to study these costs. In our paradigm, participants always perform the same task (color-naming Stroop task) but must switch between control states that best serve their current performance goal (e.g., performing the task quickly or accurately). We leverage computational models of decision-making and cognitive control to operationalize and measure control states, allowing us to study the dynamics of transitions between them. We then develop a dynamical systems model of goal switching, which assumes that the cognitive system is subject to a cost of traversing from one control state to another. This model predicts that people will settle on a “target” control state (e.g., high level of focus and caution) when continually pursuing a given performance goal (e.g., “be accurate”), and will settle on a different target control state (e.g., low level of focus and caution) when continually pursuing a different performance goal (e.g., “be fast”). Critically, the model predicts

that when transitioning between performance goals people will undershoot these target states (i.e., that they will exhibit control adjustment costs), and that these adjustment costs will specifically emerge as a function of: (1) the need to adjust control states, (2) the distance that needs to be traversed in control space, (3) the time allowed for transitions, and (4) the frequency of goal switches. We confirm these predictions across four experiments. Our findings suggest that the costs associated with switching between goals emerges from the dynamics of cognitive control adjustments which are gradual and characterized by inertia.

Control Adjustment Costs: A Dynamical Systems Model

Consider an environment in which participants perform a control-demanding task such as naming the ink color of a color word (Figure 1A). While they continue performing that task, their instruction for how to perform the task (their *performance goal*) varies: sometimes they are asked to perform the task as accurately as possible (Accuracy goal); other times they are asked to perform the as fast as possible (Speed goal). Adjusting to either goal will require participants to reconfigure their information processing accordingly. Such an adjustment can be operationalized as dynamic switch between control states (e.g., low level of caution in the speed goal vs. high level of caution in the accuracy goal). Building on previous work in switching between discrete control states (Ueltzhöffer et al., 2015; Musslick et al., 2019; Steyvers et al., 2019), we formalize this process as a dynamical system in which control is gradually adjusted from one control state (e.g., low decision threshold) to another (e.g., high decision threshold). This model's key assumption derives from the concept of task-set inertia, which has been used to explain performance costs associated with goal switches between tasks (Alport et al., 1994; Gilbert & Shallice, 2002), but has never been evaluated in the context of switching goals within a task. The intuition is that such inertia in control adjustments will produce suboptimal behavior when

people frequently switch between different control states. In order to simulate behavior arising from this model, and compare it to empirical data, we formalize the model as follows.

First (Figure 1B-left), the target control state for a given goal is determined by finding the control settings that would optimize one's reward rate (e.g., depending on the current incentives or performance goals). That target control state could be identical to or differ from the previous control state (e.g., the settings used for the most recent trial). Second (Figure 1B-middle), to the extent these states differ, the previous control state is gradually adjusted toward the target state. Third (Figure 1B-right), when the task onsets, the control state at that moment is used to determine response selection; critically, this control state that is ultimately implemented on a given trial need not match either the previous control state or the target state. We next describe each of these processing stages and model components in turn.

Determining Control Signals for Each Goal

Control states are not directly observable, and thus studying adjustments in control requires a process model that allows for the specification of control signals. In the case of task goals that pertain to speed and/or accuracy, control signals are commonly operationalized through sequential sampling models of decision-making (Ratcliff & Rouder, 1998; Bogacz et al., 2006; Ratcliff & McKoon, 2008; Ritz et al., 2022). The Drift Diffusion Model is a model of decision-making that casts decision-making as a noisy accumulation of evidence towards one of the response boundaries. The speed of evidence accumulation (*drift rate* parameter) relates to processing efficiency, which can be modulated by attention. Higher drift rates lead to faster and more accurate responses. Response caution (*threshold* parameter) represents how much evidence is needed to make a decision. Higher thresholds are related to slower, but more accurate responses.

Levels of drift rate and threshold represent two dimensions under simultaneous control within a multivariate space of possible control signals (Bogacz et al., 2006; Musslick et al., 2015; Ritz et al., 2022). Computational models of cognitive control can identify a point in this multidimensional space (a *control state*) that maximizes one's current expected benefits while minimizing control costs. To identify optimal control states for the two performance goals described above (Accuracy vs. Speed Goal), we can leverage a normative model proposed by Bogacz and colleagues (2006), that was recently further developed and empirically tested by Leng and colleagues (2021). This model finds optimal control states by maximizing the expected reward rate for different combinations of drift rates and thresholds (see the Methods section for details). We use this model to specify optimal control states (v^*, a^*) given the current state of the environment (including current performance goals; Figure 1B-left). These optimal control states can be understood as representing the target state for a given performance goal (e.g., when given the Speed goal), but how it is that a person shifts between control states remains unspecified.

Cognitive Control Dynamics

When the environment changes (e.g., a change in the explicit performance goal or in the incentives for the current goal), the current control state needs to be adjusted to a new target state in order to perform optimally on the task (Figure 1B-middle). We formalize this idea by proposing that a control state at a given time (u_t) is gradually adjusted towards the new target state (u^*). Here we consider two control signals: drift rate (v_t) and threshold (a_t) which are adjusted towards the new target states described above $(u^* \in \{v^*, a^*\}$. We represent these control states as active units in a dynamical system in which activity states correspond to the activity of control units in a neural system (cf. Ueltzhöffer et al., 2015; Musslick et al., 2018, 2019). We

simulate two control units whose activity at time t corresponds to the current level of drift rate v_t and threshold a_t. Building on previous work in task-switching (Musslick et al., 2019; Steyvers et al., 2019), the dynamics of control adjustments in this system are governed by an inertia parameter (τ) and by the distance between the current goal state (u_t) and the target goal state (u^*):

$$\frac{du}{dt} = -\tau\left(u - \frac{1}{1 + e^{-g(u^* - \theta)}}\right) + N(0, c^2) \quad (1)$$

This model can be rewritten in the following way and understood as a leaky accumulation (Usher & McClelland, 2001) of the current control state:

$$\frac{du}{dt} = -\tau u + \tau\frac{1}{1 + e^{-g(u^* - \theta)}} + N(0, c^2) \quad (2)$$

Target control state (u^*) represents the input to the system and is passed through a sigmoid activation function. The sigmoid activation function restricts the control intensity between the values of 0 and 1, and is offset by a bias term θ. The gain g regulates the sensitivity of the current processing intensity to changes in the input. This gain parameter has been suggested to modulate the impact of cognitive processes, similar to the effects observed with neuromodulatory neurotransmitters like dopamine (Servan-Schreiber et al., 1990; Liljenström, 2003; Cools, 2015), in order to control the balance between stability and flexibility in the system (Musslick et al., 2018, 2019). The inertia parameter τ determines the rate of change in the input, which is determined by two factors: (1) a decay term that scales with current control level u_t, and (2) an excitatory term that represents the optimal control level u^* for that control unit. Consequently, the equilibrium state of the control unit's net input is achieved when its control level matches the optimal control level. Noise is zero-mean Gaussian with variance set by c. The dynamics dictated above highlight two key factors that will determine the degree to which one's control

state shifts between the current and the target state: distance between the two states (e.g., how different the current and target drift rates and thresholds are) and time to adjust (e.g., the time between being presented with the performance goal and task onset).

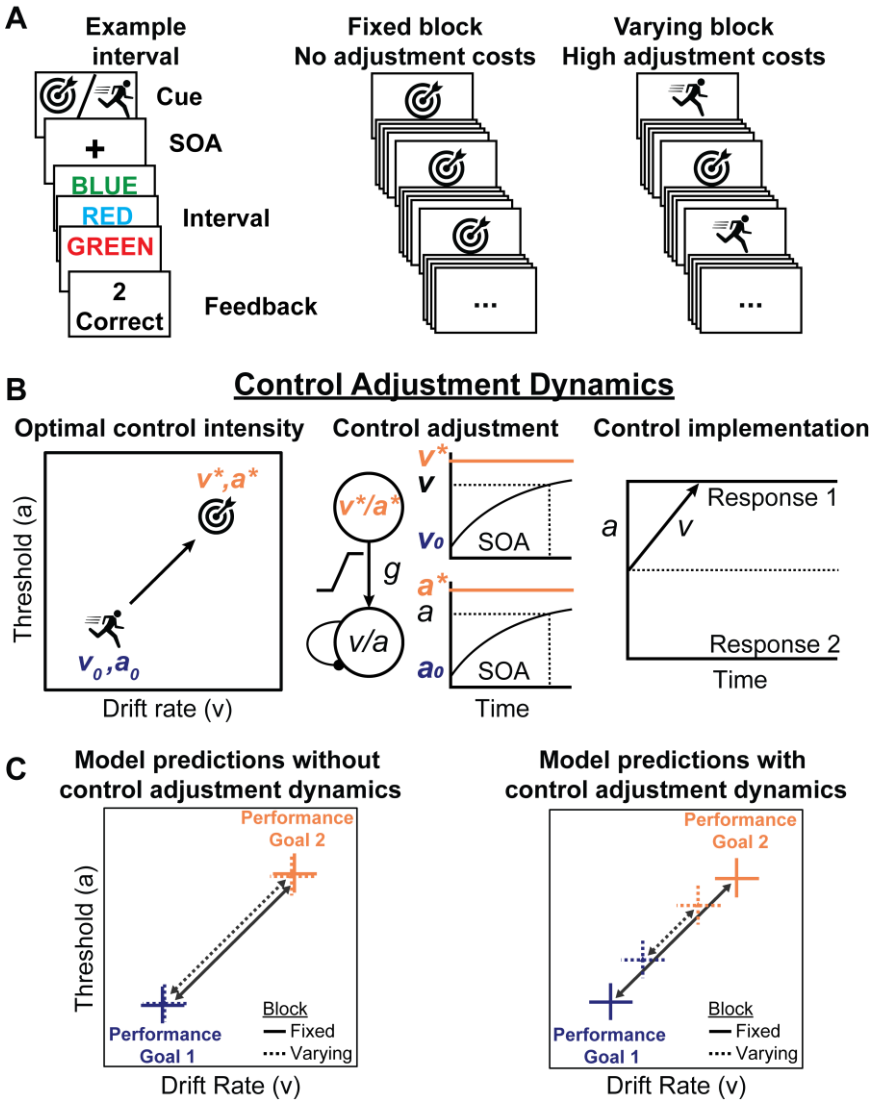

Figure 1. Behavioral task and the model of control dynamics. A. Participants performed an interval-based version of the Stroop task in which they had a fixed amount of time to complete as many trials as they wished. Prior to each interval they received saw a cue instructing them about the current performance goal (e.g., speed or accuracy). Each experiment consisted of blocks in which the performance goal never changed (fixed) and blocks in which the performance goals frequently changed (varying). **B.** After determining the optimal control levels, the model gradually adjusts drift rate and threshold from the current to the target control state. **C.** The model without such control dynamics predicts that control state will be immediately updated to the desired state. Thus, there will be no differences in implemented drift rates and threshold between situations in which control state needs to be frequently adjusted (e.g., performance goals vary) and situations in which the control state always remains the same (e.g., performance goal is always the same). The model which includes control adjustment dynamics predicts that implemented control states will be pulled closer together in blocks which involve frequent adjustments of control.

Control Implementation

Once the adjustment of control signals is terminated (e.g., task onset), the current control state is implemented and used to generate behavior (Figure 1B-right). Based on this control state (v, a) and the features of the stimulus (e.g., whether the word and ink color indicate congruent or incongruent responses), a distribution of expected responses and response time likelihoods can be estimated from the Drift Diffusion Model (Navarro & Fuss, 2009).

Model Predictions

The core assumption of the proposed model is that adjustments in control states (attention and caution) are subject to inertia (Figure 1C). Such inertia can prevent the control system from reaching a target control state. We refer to this undershoot in control adjustment as a *control adjustment cost*, and quantify this as the distance between the control state one achieves on a given trial and the inferred target control state for that trial (i.e., the control state one would settle on in the absence of a need for adjustment). This model makes a series of falsifiable predictions, which we systematically test in the studies reported below. First, relative to task environments that maintain the same performance goal (and thus no need for control adjustment), environments that require regular changes in performance goals (and thus changes in control states) should result in smaller control adjustments on average (i.e., should incur larger adjustment costs) (Study 1). Second, participants should incur less of an adjustment cost when switching between two target states that are closer to one another in control space (e.g., based on the similarity of the corresponding performance goals) and more of an adjustment cost when those target states are further apart (Study 2). Third, participants should incur more of an adjustment cost the less time they are given to transition between their current control state and a different target state (e.g., based on the relative onset of goal and task stimulus information)

(Study 3). Finally, participants should incur greater control adjustment costs the more frequently they are required to transition between control states (Study 4).

Results

Adjustment Costs Arise Under Conditions of Changing Performance Goals (Study 1)

Participants in Study 1 (N = 44) performed the Stroop task during fixed time intervals (8-12s). On each interval, they were instructed to perform the task as fast (Speed goal) or as accurately (Accuracy goal) as possible. They did this in blocks in which the performance goal was either always the same (Fixed blocks), or the two goals alternated (Varying blocks).

Model Predictions

Our model predicts that there is inertia in the adjustments of control signal, which will lead to different behavioral profiles in situations in which participants always have the same control signals compared to situations in which they have to frequently adjust control (Figure 1C). Thus, the model predicts that control intensities for the speed and accuracy conditions will differ between fixed and varying blocks, and that these will manifest in changes to drift rates and thresholds. To formalize this prediction, we first simulated what level of drift rates and thresholds participants would set in the speed and accuracy conditions if there were no costs to switching between them (i.e., if one of them was their only goal; details in Methods section). We proceeded to simulate which levels of drift rate and threshold will be reached on average in blocks in which participants had to frequently switch between different goals (varying blocks). These simulations assumed that participants tried to reach the same levels of drift rates and thresholds as in fixed blocks, but failed to do so because they had to frequently adjust control states. Thus, control adjustment costs emerge, and they are revealed by the prediction that the distance between two control states is reduced in varying compared to fixed blocks (Figure 2B-

left). This effect is generated by the need to frequently adjust control signals (move in the 2-D control space) in the varying blocks. Because of the gradual adjustment of control assumed by our model, such movement means that the control system will often be in between the two target states, rather than at the target state required by the current goal. Thus, the second prediction of our model (Figure 2D-left) is that in varying blocks control intensities for the two goals will be closer together on switch intervals (in which the agent has a different goal state compared to the previous interval) relative to the repeat intervals (in which the agent has the same goal state as on the previous interval). We tested these two findings by first comparing the behavioral performance of our participants on fixed and varying blocks, as well as on switch and repeat intervals within the varying blocks. Then, we fit the Drift Diffusion Model to behavioral data in order to obtain the drift rates and thresholds across conditions and compare them to model predictions. Across these analyses, the key index of control adjustment costs is the extent to which people adjust their control state less between the two performance goals when they are in varying relative to fixed blocks. Control adjustment costs are evidenced by decreased difference between Speed and Accuracy conditions when participants frequently switch between these two performance goals. We tested for these costs based on differences in raw behavioral performance across the conditions (accuracies and RTs; *model-agnostic* indices of control adjustment costs) and based on control state estimates from fitting behavioral data to a Drift Diffusion Model (*model-based* indices of control adjustment costs).

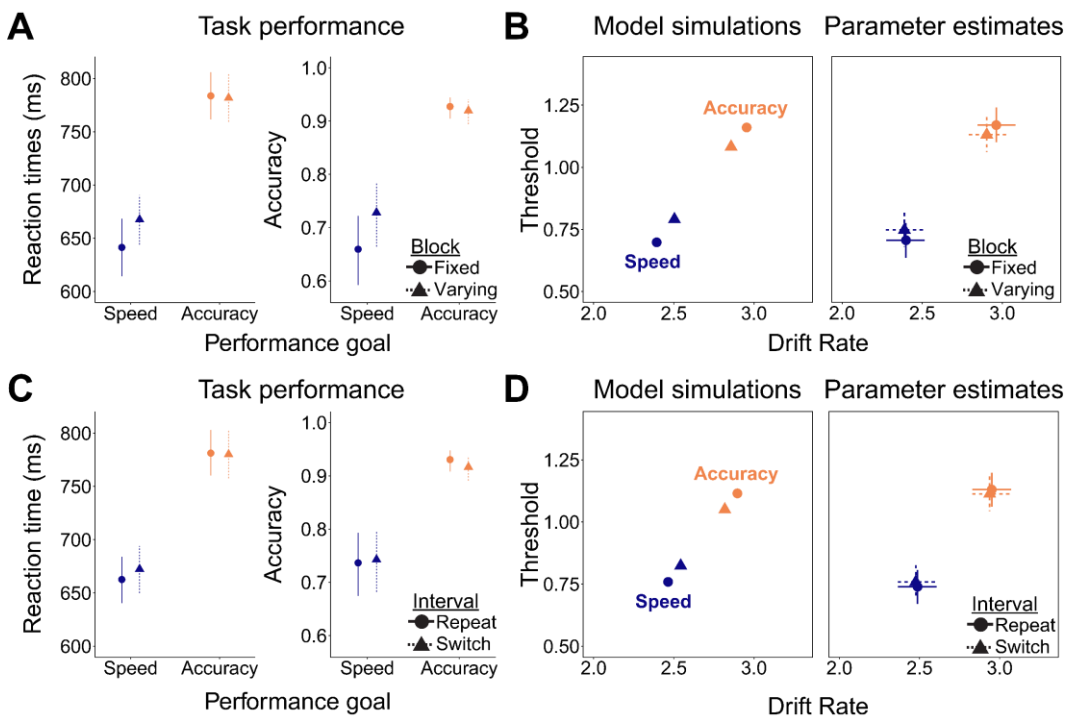

Figure 2: Frequent adjustments of control intensity impact behavioral performance and control intensity. A. In the speed condition (blue) participants were faster and less accurate in blocks with the fixed goal (solid) compared to blocks in which the goals were varying (dashed). Dots represent regression estimates and error bars 95% credible intervals. **B.** Model simulations (left) and drift diffusion parameter estimates (right) show that control intensities for the two performance goals are pulled closer in varying blocks which include goal switches. Error bars represent 95% credible intervals. **C.** In varying blocks, participants were faster in the speed condition when the previous interval had the speed (Repeat) relative to accuracy (Switch) goal. In the accuracy condition, they were less accurate in the switch relative to speed condition. **D.** Model simulations and parameter estimates for the switch and repeat intervals in varying blocks show that the estimates are pulled closer together on switch relative to repeat intervals.

Model-Agnostic Evidence of Control Adjustment Costs

As shown in Figure 2A and Table S1, participants were slower (b=128.25; 95% CrI [85.36, 172.52]; $p_{b<0}$=0.00) and more accurate[1] (b=1.67; 95% CrI [1.12, 2.22]; $p_{b<0}$=0.00) when their performance goal was to maximize accuracy compared to when it was to maximize speed. Crucially, we found evidence for control adjustment costs when comparing fixed and varying

[1] All of the estimates for accuracy analyses are in log odds.

blocks. The difference between speed and accuracy conditions was larger in fixed compared to varying blocks for both reaction times (b=28.09; 95% CrI [3.54, 51.41]; $p_{b<0}$=0.01) and accuracy (b=0.44; 95% CrI [0.12, 0.76]; $p_{b<0}$=0.00). These results indicate that the performance in the speed condition was closer to the performance in the accuracy condition when participants had to switch between these strategies within one block relative to when they had only one performance goal per block. We confirmed (Table S2) that these findings are not driven by changes in vigor (non-decision time) nor the reduction of thresholds over time (collapsing bound).

We next analyzed behavior in the varying block, focusing our analysis on whether the performance goal on the previous interval was the same as on the current one (repeat), or different than the current goal (switch). This analysis (Figure 2C) revealed that the difference between accuracy and speed performance goals (reaction times: b=113.19; 95% CrI [71.51, 156.99]; $p_{b<0}$=0.01; accuracy: b=1.45; 95% CrI [0.87, 2.02]; $p_{b<0}$=0.01), was reduced when the current interval included a goal switch compared to a goal repeat (reaction times: b=10.97; 95% CrI [-1.65, 23.86]; $p_{b<0}$=0.04; accuracy: b=0.23; 95% CrI [0.01, 0.46]; $p_{b<0}$=0.02).

Model-based Evidence of Control Adjustment Costs

Our model predicts that, due to control adjustment costs, drift rates and thresholds for the two performance goals will be closer together in blocks in which the goals vary compared to blocks in which they are fixed (Figure 2B-left). To test this prediction, we fitted a drift diffusion model to behavioral data in order to jointly model both the reaction times and accuracies, and to directly probe the prediction from our model. This analysis revealed a substantial difference in both drift rates and thresholds between the speed and accuracy performance goals. Participants had higher drift rates (b=0.54; 95% CrI [0.31, 0.77]; $p_{b<0}$=0.00) and thresholds (b=0.42; 95% CrI

[0.27, 0.58]; $p_{b<0}$=0.00) when they were instructed to focus on accuracy, relative to when they were instructed to focus on speed (Figure 2B-right, Table S1, and Figure 6).

As predicted by our model (Figure 2B-left), the distance between the speed and accuracy conditions was reduced in varying compared to fixed blocks. This was true in thresholds (b=0.08; 95% CrI [0.03, 0.13]; $p_{b<0}$=0.00; Figure 2B-right), and held in the same direction for drift rates, but not reliably so (b=0.05; 95% CrI [-0.11, 0.21]; $p_{b<0}$=0.28).

The second prediction from our model is that the distance between drift rates and thresholds among the speed and accuracy conditions should be reduced on switch compared to repeat intervals (Figure 2D-left and Table S3). To test this prediction, we focused our analysis on the varying blocks and compared the difference between accuracy and speed intervals in cases which required a switch to the cases which did not require switches (i.e., repeats of the performance goal). This analysis revealed that the effect of performance goal on thresholds was larger in repeat relative to switch intervals (b=0.04; 95% CrI [0.00, 0.07]; $p_{b<0}$=0.02). Similar to the block-wise analysis, this difference was not found for the drift rates (b=0.00; 95% CrI [-0.15, 0.15]; $p_{b<0}$=0.49). These findings closely follow the block-wise analyses, suggesting that the differences in control intensities between fixed and varying blocks arise due to the control adjustment costs stemming from the need to frequently adjust control signals in response to changing performance goals.

Discussion

The results of Study 1 confirmed that we can induce reliable differences in behavioral performance and in estimates of underlying control signal intensities (drift rates and thresholds). This allowed us to test for the existence of control adjustment costs predicted by our model. Comparing environments which demand frequent changes in goals states to the ones which

don't, we found evidence for control adjustment costs in both behavioral performance and control intensity. Confirming the first prediction of our model, we found that control adjustment costs arise due to the demand to switch between different performance goals. Next, we tested the prediction of our model that control adjustment costs are governed by the distance between control states, with longer distances producing larger control adjustment costs.

The Costs of Control Adjustment Scale with the Distance Between Control States (Study 2)

Our model assumes that control adjustment costs depend on the distance between the current state of the control system and the desired state induced by the new performance goal. The model assumes that control intensities are gradually adjusted toward the desired state. A key prediction that follows is that the distance between the current and desired state determines control adjustment costs. We should therefore observe greater adjustment costs (as indexed by the distance between target control levels and implemented control levels) the further the current goal state is from the previous one. Study 2 tested this prediction by having participants (N = 38) perform the Stroop task under three performance goals: respond as quickly as possible (speed goal), as accurately as possible (accuracy goal), or both as fast and accurately as possible (speed/accuracy goal). Participants performed the task in blocks in which the goal was always the same (fixed blocks), or in blocks in which the three goals frequently changed (varying blocks). This paradigm allowed us to investigate whether control adjustment costs depend on the distances between different target goal states.

Model Predictions

Simulations of this task were generated in the same way as in Study 1. We first simulated fixed blocks in which the agent performed under only one of the three performance goals (speed,

accuracy, and speed/accuracy). We then compared performance in fixed blocks to simulated performance in varying blocks in which the agent switched between the three performance goals. The model predicted (Figure 3B-left) larger control adjustment costs for the conditions that were further in control state space (speed vs. accuracy), relative to the conditions which were closer (accuracy vs. speed/accuracy).

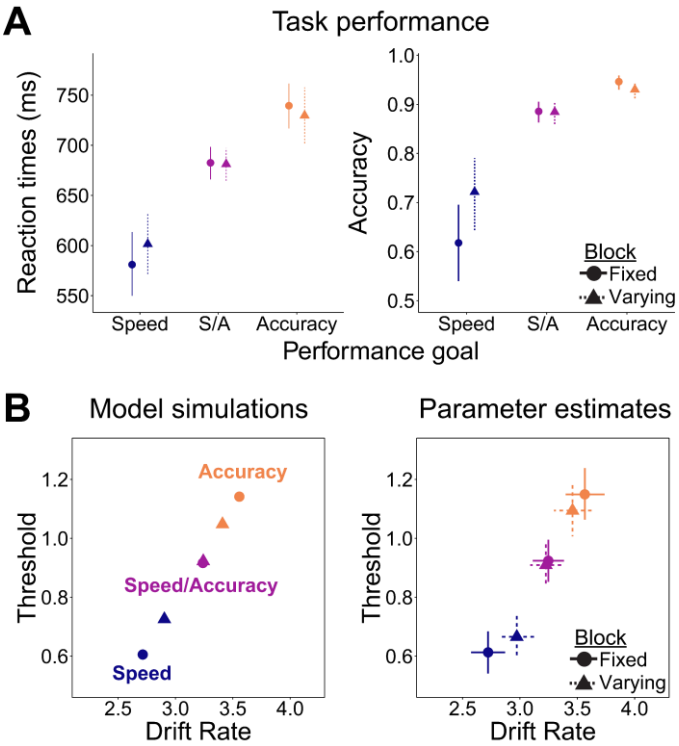

Figure 3. Magnitude of control adjustments impacts behavioral performance and control intensity. A. Reaction times and accuracies for the three performance goals in fixed and varying blocks. Dots represent regression estimates and error bars 95% credible intervals. **B.** Model simulations (left) and drift diffusion parameter estimates (right) for the three performance goals. Error bars represent 95% credible intervals.

Model-Agnostic Evidence of Control Adjustment Costs

Replicating the first experiment, we found that participants were slower (b=143.32; 95% CrI [95.18, 190.48]; $p_{b<0}$=0.00) and more accurate (b=2.01; 95% CrI [1.50, 2.52]; $p_{b<0}$=0.00) when they performed under the accuracy compared to the speed goal (Figure 3A and Table S4). In the speed/accuracy condition, participants' performance was in between that of the other two

performance goals. Relative to the accuracy condition, participants were faster (b=-52.74; 95% CrI [-79.68, -26.12]; $p_{b>0}$=0.00), but less accurate (b=-0.69; 95% CrI [-79.10, -25.10]; $p_{b>0}$=0.00). Thus, these three performance goals induced distinct behavioral performance profiles.

Replicating Study 1, we found evidence of control adjustment cost when comparing speed and accuracy conditions in varying and fixed blocks. The costs were present in both reaction times (b=30.30; 95% CrI [-3.83, 64.80]; $p_{b<0}$=0.04) and accuracy (b=0.75; 95% CrI [0.31, 1.20]; $p_{b<0}$=0.00). Critically, this control adjustment costs (i.e., performance difference between performance goals across fixed and varying blocks) was not found when comparing other performance goals which were closer together, thus requiring less adjustment. We found no evidence of adjustment costs between accuracy and speed/accuracy conditions (reaction times: b=-8.43; 95% CrI [-35.40, 17.49]; $p_{b>0}$=0.26; accuracy: b=-0.27; 95% CrI [-0.64, 0.10]; $p_{b>0}$=0.08), nor between speed and speed/accuracy (reaction times: b=-26.65; 95% CrI [-61.21, 8.64]; $p_{b>0}$=0.10; accuracy: b=-0.07; 95% CrI [-0.39, 0.25]; $p_{b>0}$=0.36).

Model-based Evidence of Control Adjustment Costs

During fixed blocks, participants had the highest drift rates and thresholds in the accuracy condition, and lowest in the speed condition (Figure 3B-right). The speed/accuracy condition was in between the other two. We found significant differences between both accuracy and speed/accuracy conditions (drift rates: b=0.27; 95% CrI [0.09, 0.46]; $p_{b<0}$=0.01; thresholds: b=0.20; 95% CrI [0.10, 0.31]; $p_{b<0}$=0.00) and between speed/accuracy and speed conditions (drift rates: b=0.39; 95% CrI [0.27, 0.51]; $p_{b<0}$=0.00; thresholds: b=0.28; 95% CrI [0.20, 0.35]; $p_{b<0}$=0.00). The distance between the speed and the speed/accuracy condition was larger than the distance between the accuracy and the speed/accuracy condition, but this difference was not robust (drift rates: b=0.21; 95% CrI [-0.10, 0.51]; $p_{b<0}$=0.13; thresholds: b=0.08; 95% CrI [-0.07,

0.24]; $p_{b<0}$=0.18). Thus, the three performance goals successfully produced different control states. This allowed us to test the prediction from our model that conditions farther away in control space (speed vs. accuracy) should result in higher control adjustment costs than conditions closer together (speed/accuracy vs. accuracy and speed/accuracy vs. speed).

Replicating Study 1, we found evidence of control adjustment costs when comparing performance between the accuracy and speed goals (*b*=0.36; 95% CrI [0.16, 0.55]; $p_{b<0}$=0.00) and thresholds (*b*=0.11; 95% CrI [0.04, 0.18]; $p_{b<0}$=0.00; Figure 3B-right, Table S4, and Figure 6). Of note, in this experiment we found adjustment costs in both threshold and in drift rate (which was not significant in Study 1). Critically, we confirmed the second prediction of our model, as we found no evidence of control adjustment costs between two closest conditions – accuracy and speed/accuracy – in neither drift rates (*b*=0.09; 95% CrI [-0.09, 0.27]; $p_{b<0}$=0.21), nor thresholds (*b*=0.04; 95% CrI [-0.03, 0.11]; $p_{b<0}$=0.17). However, we did find significant adjustment costs, though smaller in magnitude, when comparing the speed/accuracy to the speed condition (drift rates: *b*=0.27; 95% CrI [0.03, 0.51]; $p_{b<0}$=0.03; thresholds: *b*=0.07; 95% CrI [0.02, 0.12]; $p_{b<0}$=0.01). Thus, we found that distance in control space determines control adjustment costs, with increased costs in situations which demand larger adjustments in control.

Discussion

Results of Study 2 confirmed our model's prediction that control adjustment costs depend on the distance between the current and the desired goal state. By investigating behavioral performance and control signal intensities across three performance goals we showed that the costs increase as a function of the distance between the goal-optimal points in the control state space. We next tested the prediction that control states are gradually adjusted over time. We thus tested the impact of time allowed to change control signals on control adjustment costs.

The Costs of Control Adjustments Depend on the Time Allowed for Adjustments (Study 3)

Our model assumes that control signals are adjusted incrementally from the current to the desired control state. To investigate the role of time in control adjustments, in this study we had participants (N=50) perform the same task as in Study 1 (speed vs. accuracy goals in fixed and varying blocks), but we varied the amount of time they had to adjust control signals. In half of the blocks participants had a shorter amount of time to adjust control (250ms stimulus-onset asynchrony (SOA) between the goal cue and task onset), while in the other half of blocks they had more time (1000ms SOA) to adjust control. In addition, in this experiment we made goal switches fully predictable (see Methods for details) to ensure that control adjustment costs are not dependent on whether on uncertainty about the current or the next goal.

Model Predictions

Simulations of this study were the same as in Study 1, but we now also simulated blocks with shorter or longer SOA by having the model take either 250 or 1000 steps before implementing the reached control state. The model predicts that shorter adjustment windows will result in increased control adjustment costs because there is less time to move toward the desired control state before task onset (Figure 4B-left).

Model-Agnostic Evidence of Control Adjustment Costs

Replicating our previous studies, we found that participants were slower (b=50.66; 95% CrI [29.08, 71.58]; $p_{b<0}$=0.00) and more accurate (b=0.88; 95% CrI [0.57, 1.20]; $p_{b<0}$=0.00) when instructed to focus on accuracy relative to speed (Figure 4A and Table S5). We again found evidence for control adjustment costs when comparing performance in varying and fixed blocks (reaction times: b=14.37; 95% CrI [-1.68, 29.37]; $p_{b<0}$=0.04; accuracy: b=0.62; 95% CrI [0.36, 0.87]; $p_{b<0}$=0.00). We found suggestive evidence indicating that control adjustment costs were

larger in blocks in which people had shorter time to adjust (SOA =250ms) compared to when they had more time (SOA=1000ms) in both reaction times (b=-19.55; 95% CrI [-51.10, 11.65]; $p_{b>0}$=0.10) and accuracy (b=-0.20; 95% CrI [-0.61, 0.19]; $p_{b>0}$=0.16). While these interactions were not individually significant, both RT and accuracy effects were directionally consistent with the expected direction of change, suggesting that they may jointly reflect a reliable change in the relevant model-based indices of control costs.

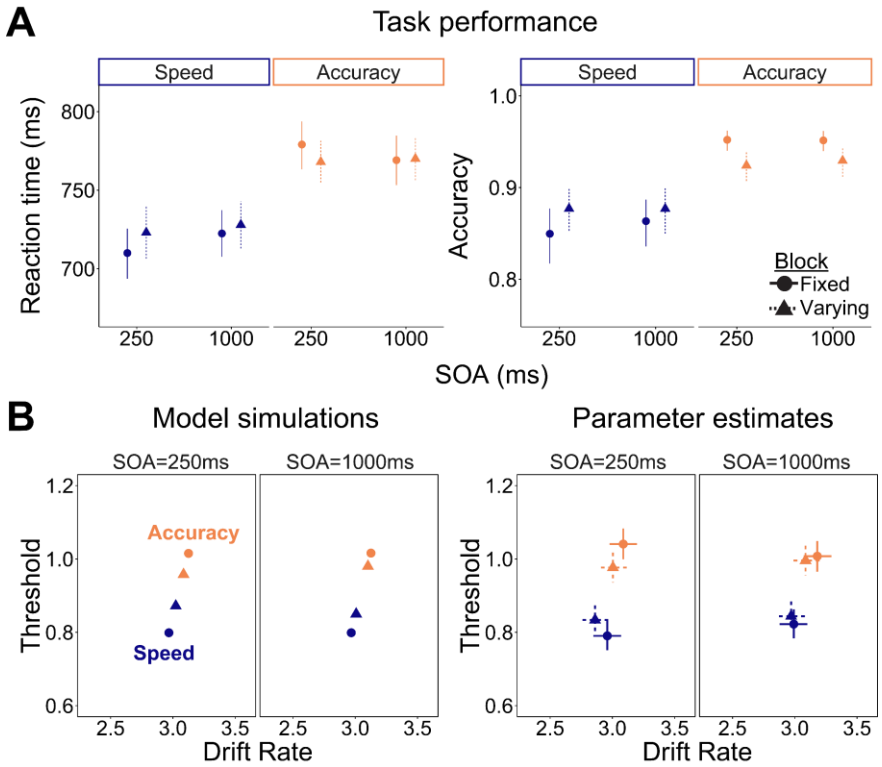

Figure 4. Control adjustment costs are increased when there is less time to change control signals. A. Reaction times and accuracies for the speed and accuracy performance goals when the time to adjust control (SOA between goal cue and task onset) is 250 vs. 1000ms. Dots represent regression estimates and error bars 95% credible intervals. **B.** Model simulations (left) and drift diffusion parameter estimates (right) speed and accuracy performance goals when SOA is 250 vs. 1000ms. Error bars represent 95% credible intervals.

Model-based Evidence of Control Adjustment Costs

Replicating our previous results, we found (Figure 4B-right, Table S5, and Figure 6) that drift rates (b=0.14; 95% CrI [0.06, 0.23]; $p_{b<0}$=0.00) and thresholds (b=0.18; 95% CrI [0.12, 0.24]; $p_{b<0}$=0.00) were higher in the accuracy relative to speed condition. We also found

evidence for control adjustment costs in threshold (b=0.07; 95% CrI [0.03, 0.11]; $p_{b>0}$=0.00). Critically, control adjustment costs were reduced when participants had more time to adjust control (b=0.07; 95% CrI [0.00, 0.15]; $p_{b>0}$=0.03). For drift rates, we found no evidence of control adjustment cost, nor a change in the cost due to the amount of time to adjust control (ps>0.3). These results confirm our model's prediction that suboptimal adjustments can arise from insufficient time to transition between current and target levels of control (Figure 4B-left).

Discussion

In this study we demonstrated that control adjustment costs depend on the time allowed for the adjustment. Since our model assumes that control states are gradually adjusted from the current toward the desired state, the model predicts decreases in adjustment costs as the time allowed for the adjustment increases. Our findings confirm this prediction, showing that control adjustment costs were higher when participants had less time to adjust control. Next, we tested the prediction that control adjustment costs depend on the frequency of goal switches, with more frequent switches producing larger adjustment costs.

Control Adjustment Costs Depend on Goal Switch Frequency (Study 4)

Our model predicts that the number of goal state switches determines the control adjustment costs. In this experiment, we directly tested this prediction by comparing fixed blocks (no goal switches) to varying blocks (randomly in Studies 1 and 2 and every interval in Study 3). Further, we explicitly instructed participants about the goal switch frequency prior to the start of each block (cf. Liu & Yeung, 2020). Thus, we could investigate the effect of goal switch frequency on control adjustment costs. Participants (N=55) performed the same Stroop task

under speed and accuracy goals. Again, we had fixed blocks in which the goal never changed. In three other block types, goals predictably switched every 8, 4, or 2 intervals.

Model Predictions

To generate model predictions for this study, we simulated the conditions in which the speed and accuracy performance goals never change (fixed blocks), as well as the blocks in which the goals changed every 8, 4, and 2 intervals (Figure 5B). Importantly, unlike these previous studies, participants in this study were provided with advance information about goal switch frequency in the upcoming block. Previous work on task-switching has shown that frequent switches demand higher flexibility, which can be captured in models with changing how sensitive the model is to the input about the current task (Ueltzhöffer et al., 2015; Musslick et al., 2018, 2019). Higher sensitivity to such input leads to faster task-switches, at the cost of performance cost of each task. To implement this in our model we adjusted the gain parameter (g in Equation 1) of the sigmoid activation function across the 4 block types (see Methods for details). Lower values of gain correspond to more stable control configurations, and thus, are optimal in environments in which task goals switch more frequently (Musslick et al., 2018). Thus, g can be thought of as a meta-parameter, that may be set by the participant depending on their expectation about the volatility of the environment. Following previous work (Musslick et al., 2018, 2019), we assumed that participants could proactively adjust gain (Figure 5B). We parametrically decreased gain across blocks with increasing number of goal switches. This allowed for more rapid goal switches in blocks in which participants had to frequently adjust control signals. Both the model with and without gain adjustment predict increased control adjustment costs with increased frequency of goals switches. However, the model which includes a gain change predicts a much larger change in control adjustment costs because it assumes that

participants proactively adjust how flexible they are based on the knowledge about the frequency of goal switches.

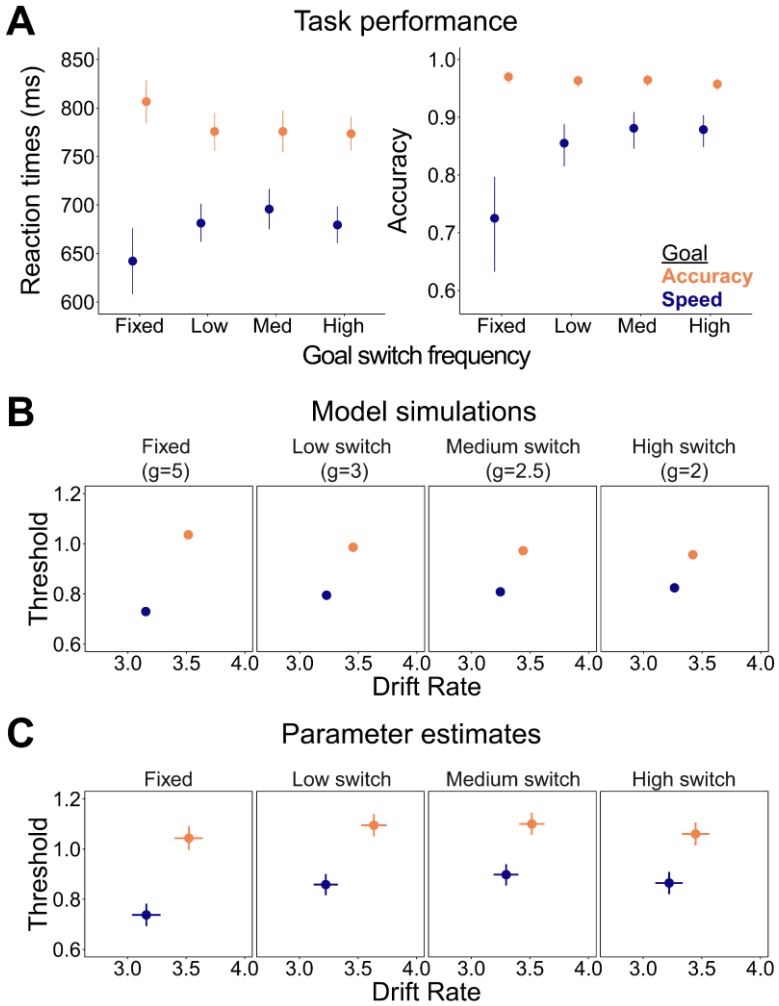

Figure 5. Control adjustment costs increase with the number of performance goal switches. A. Reaction times and accuracies across blocks in which performance goals are fixed, or they change every 8 (low), 4 (medium), or 2 (high) intervals. Dots represent regression estimates and error bars 95% credible intervals. **B.** Model simulations across four levels of goal switch frequency. **C.** Drift diffusion parameter estimates. Error bars represent 95% credible intervals.

Model-Agnostic Evidence of Control Adjustment Costs

Participants were once again slower (b=54.14; 95% CrI [37.96, 70.30]; $p_{b<0}$=0.00) and more accurate (b=0.80; 95% CrI [0.58, 1.04]; $p_{b<0}$=0.00) when instructed to focus on accuracy compared to speed (Figure 5A). Critically, control adjustment costs increased with the increasing frequency of goal switches (Figure 5A and Table S6). This was true for both the reaction times

(b=25.20; 95% CrI [8.50, 42.58]; $p_{b<0}$=0.00) and accuracy (b=0.48; 95% CrI [0.25, 0.74]; $p_{b<0}$=0.00). This result suggested that control adjustment costs depend on the number of times participants had to adjust control intensity.

Model-based Evidence of Control Adjustment Costs

Replicating our previous studies, we found higher drift rates (b=0.30; 95% CrI [0.17, 0.44]; $p_{b<0}$=0.00) and thresholds (b=0.23; 95% CrI [0.17, 0.29]; $p_{b<0}$=0.00) in accuracy compared to speed intervals. Importantly, as predicted, control adjustment costs increased with more goals switches in a block (Figure 5C, Table S6, and Figure 6), for both drift rates and thresholds (interaction with the linear effect of block type: drift rates: b=0.20; 95% CrI [0.05, 0.43]; $p_{b>0}$=0.06; thresholds: b=0.12; 95% CrI [0.04, 0.21]; $p_{b<0}$=0.00). Thus, control adjustment costs varied with the expected frequency of goal switches. This result provided support for the model which assumed that participants change the overall level of flexibility depending on the instructed frequency of goal switches (Figure 5B and Figure S1).

Discussion

In this study we confirmed that control adjustment costs depend on the frequency with which participants have to adjust their control states. Critically, we show that participants proactively adjust how flexible they are across blocks which demand different number of goal switches. We showed that the control adjustment costs parametrically increase as the number of goal switches within a block increase. This effect was captured by the model which assumed that the expectation of more or less goal switches leads to a change in the gain parameter which allows for more rapid switching, but higher control adjustment cost, in blocks with high relative to low goal switch frequency.

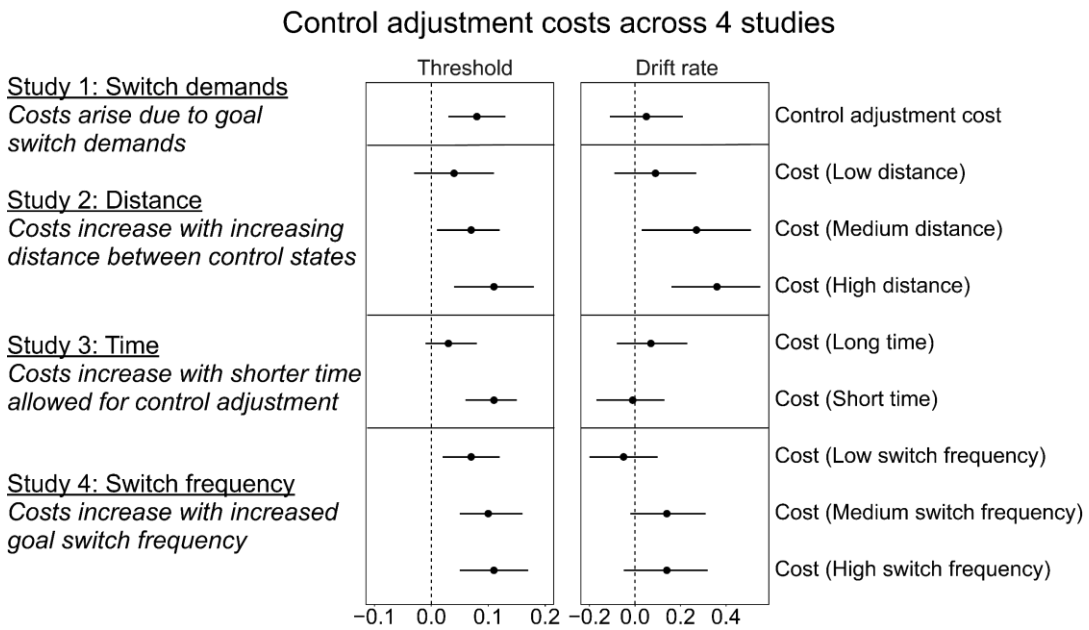

Figure 6. Summary of model-based control adjustment costs and their causes across four studies. Control adjustment costs are indexed by the extent to which control states corresponding to different performance goals are adjusted less in blocks in which performance goals vary relative to fixed blocks. Points represent Drift Diffusion Model estimates, and error bars are 95% credible intervals. Study 1 shows a significant adjustment cost on threshold. Study 2 shows the smallest cost between states with low distance (Accuracy vs. Speed/Accuracy goal), larger cost with medium distance (Speed/Accuracy vs. Speed), and largest cost with high distance (Accuracy vs. Speed goal). Study 3 shows a larger cost in threshold when people have less time to adjust control states. Study 4 shows that relative to the fixed block, the cost is smallest when people switch goals every 8 intervals, larger when they switch every 4 intervals, and largest when they switch every 2 intervals.

General Discussion

One of the hallmarks of human behavior is the ability to flexibly adjust to the demands of the current environment. Frequent changes in our own goals, as well as in our environments, demand adjustments in our behavior. These adjustments are supported by changing cognitive control states, but such changes incur costs. Previous work has focused on the costs associated with adjustments in which stimuli are processed, and which actions need to be taken in response (Alport et al., 1994; Rogers & Monsell, 1995; Monsell, 2003; Kiesel et al., 2010). Here we extend the understanding of the costs of switching between goals by focusing on the costs which arise from the need to adjust how information is processed and how responses are selected,

which have not been studied before. We show that such adjustment produces robust costs in situations which demand performance goals changes within a single task. We isolate these control adjustment costs in a novel behavioral paradigm and test predictions of a computational model showing that they can arise due to the dynamics of changes in cognitive control states that support goal-directed behavior.

Here we focus on adjustments of continuous control states defined by the amount of attention and caution required by the current goal. Such states broaden the notion of a task-set , which commonly specifies a control state with respect to which stimuli are attended and which responses are taken (Alport et al., 1994; Rogers & Monsell, 1995). Mounting empirical evidence shows that people adjust their performance within a task based on performance goals (Forstmann et al., 2008; Ratcliff & Rouder, 1998) and expected incentives (Manohar et al., 2015; Parro et al., 2018; Frömer et al., 2021). The Drift Diffusion model (Ratcliff & McKoon, 2008; Bogacz et al., 2006; Ritz et al., 2022) is commonly used to measure changes in as a function of instructions or incentives (Forstmann et al., 2008; Leng et al., 2021; Grahek et al., 2023). Here we leverage this model to conceptualize control states along a two-dimensional space, defined by levels of attention and caution demanded by the current goal (e.g., try to perform a task quickly or accurately). This allowed us to study continuous adjustments of cognitive control states, defined by *how efficiently* people processed stimuli (drift rate parameter, a putative analog of top-down attention) and *how cautious* they are when they act (threshold parameter). We showed that costs arise from the requirement to transition between the target control state for one goal (e.g., low attention and low caution) and the target control state for a different goal (e.g., high attention and high caution) (Study 1). We show that such costs arise in particular from a combination of the distance one needs to traverse between control states (Study 2) and the time allowed for that

transition process (Study 3), with greater costs (smaller adjustments) emerging when target control states are far apart (e.g., Speed goal → Accuracy goal) and/or transition time is short (e.g., 250ms), relative to when those control states are closer (e.g., Speed → Speed & Accuracy) and the transition time is longer (e.g., 1000ms). and subject to inertia. We show that these costs also increase as the goal switches become more frequent, and that people are able to proactively use the information about goals switch frequency to adjust how flexible they are within an environment (Study 4). This work offers a mechanistic account of the causes of costs associated with adjusting cognitive control states.

Computational models of cognitive control posit that people adapt their allocation of control to a task (including levels of attention and caution) based on the expected benefits and costs of doing so. Across different models, optimal control states are found by maximizing the expected utility of control signals (Bogacz et al., 2006; Simen et al., 2006; Shenhav et al., 2013; Manohar et al., 2015; Verguts et al., 2015; Alexander et al., 2018). In support of these models, empirical work shows that people adjust control according to their current task goal (e.g., focus on speed vs. accuracy; Forstmann et al., 2008; Ratcliff & Rouder, 1998), expected rewards (Padmala & Pessoa, 2011; Krebs et al., 2012) and penalties (Leng et al., 2021), and based on the expected costs of control (Kool et al., 2010; Westbrook et al., 2013, 2020). Thus, these models assume that levels of control depend on the incentives in one's environment, and that these will be adjusted once the incentives change (either extrinsic incentives or goals). While some of the computational models incorporate a cost for adjusting control from one setting to another (Musslick et al., 2015; Lieder et al., 2018; Grahek et al., 2020), such costs were until now only stipulated and not empirically demonstrated. Our work provides empirical support for these models, but also goes further by specifying how these costs emerge.

Investing higher levels of control (e.g., allocating more top-down attention to task-relevant stimuli) and switching between goals both incur costs. While the causes of both sets of costs remain deeply mysterious (Shenhav et al., 2017), an intriguing possibility is that these two costs are linked, with the latter in part explaining the former. Specifically, drawing on past research on the stability-flexibility dilemma – the observed trade-off between being able to focus cognitive resources on the task at hand versus rapidly re-adjust information processing (Goschke, 2013) – Musslick & Cohen (2021) propose that the control system may have a bias against maximizing its allocation of control resource to a given task because this would make it more difficult to rapidly switch to a new task. Their account predicts that the allocation of control towards a given task depends in part on the expectation about the frequency of task switches in the current environment (Musslick et al., 2019). Our theoretical and empirical findings give substance to this account. First, we propose that control adjustment costs arise from the time it takes to adjust control states from their current to the reward-optimal states. Second, we show that our model can be cast as a neural network, and that the balance between stability and flexibility can be set by adjusting the gain of the activation function (cf., Ueltzhöffer et al., 2015). We show that changes in the gain of the activation function that correspond to the expected frequency of control state switches are able to capture our empirical data well (Study 4). Finally, and of particular note, our model is able to capture these control adjustment costs without explicitly incorporating them into the evaluation that determines control allocation (an assumption that has been incorporated into some previous models in order to account for switch costs; Musslick et al., 2015; Lieder et al., 2018; Grahek et al., 2020). Rather, we propose that control adjustment costs arise from the time it takes to adjust control states, and the distance between the current and the desired control state. This proposal aligns well with the recent focus

on the dynamics of control (Kikumoto & Mayr, 2020; Steyvers et al., 2019; Holroyd, 2023), as well as with proposals that cognitive effort scales with how hard it is to reach the desired control states from the current states (Kim et al., 2018; Weninger et al., 2022).

The current study prompts a critical inquiry: do the costs tied to switching control states within a task derive from the same mechanisms postulated for inter-task switches? Some researchers advocate that task switch costs emanate from an active task-set reconfiguration process, underpinned by a control mechanism (Rogers & Monsell, 1995; Meiran, 1996; Mayr & Kliegl, 2000). In contrast, others propose these costs stem from passive processes such as proactive interference — also termed "task-set inertia" (Alport et al., 1994), inhibition of the last performed task-set (Mayr & Keele, 2000; Altmann, 2007), repetition priming of the task cue (Logan & Bundesen, 2003), or repetition priming of stimulus attributes (Wylie & Allport, 2000; Waszak et al., 2004). Our findings align with the passive process theory, ascribing within-task switch costs to the inertia from reconfiguring control states. Nevertheless, these results introduce further queries, notably whether within-task control reconfiguration costs can be tied to other passive mechanisms like repetition priming, or perhaps to active control state reconfiguration processes.

Our ability to flexibly adjust control processes lies at the core of goal-directed behavior. Given the importance of such behavior, the question of why there are costs to adjusting cognitive control state which support it, remains a puzzle within cognitive science (Musslick & Masís, 2023). By focusing on adjustments of continuous control states, rather than changes in discrete task-sets, we show how costs of adjusting control can arise from the dynamics of cognitive control. Many real-world tasks include such adjustments in attention and caution, but this aspect of control adjustments is often overlooked. Thus, studying control adjustment costs opens an

avenue for a more precise understanding of the mechanisms which lead to behavioral inflexibility, which is critical for understanding both psychopathology (Kashdan & Rottenberg, 2010) and aging (Wasylyshyn et al., 2011). In addition, our work offers a more complete understanding of incentive-based adjustments of cognitive control, which is often studied in paradigms that require frequent adjustments of attention and caution (Botvinick & Braver, 2015; Parro et al., 2018). In this way, our work adds to the growing interest in cognitive control dynamics and offers a route towards a unified view of the costs of cognitive control adjustments, which arise from the need to gradually adjust control, while facing the tradeoff between rapid adjustments and good task performance.

Methods

Participants

Across all four studies, we recruited participants with normal or corrected-to-normal vision from Prolific. All participants were fluent English speakers residing in the United States. Research protocols for all four studies were approved by Brown University's Institutional Review Board. All of the studies took approximately 1h to complete, and participants were compensated with a fixed rate of $8 per hour. The tasks never included any performance-based monetary incentives. In Study 1 we recruited 48 participants, 4 of which were excluded due to failed attention checks, yielding the final sample of 44 participants (27 female, 14 male, 3 gender not reported; median age = 31). Study 2 included 45 participants, 7 of which were excluded due to failed attention checks, and the final sample consisted of 38 participants (21 female, 17 male; median age = 36). In Study 3 we had 52 participants, 2 of which failed the attention checks, yielding the final sample of 50 participants (17 female, 33 male; median age = 33). Study 4 had

62 participants, 7 of which failed the attention check. The final sample included 55 subjects (17 female, 37 male, 1 gender not reported; median age = 31).

Task Design

Study 1

Participants performed a Stroop task in which they identified the ink color of a color word (e.g., “BLUE”) by pressing one of the 4 keys corresponding to one of the 4 colors (blue, red, yellow, or green). Half of the Stroop trials were either congruent (ink color and the color word are the same) and the other half were incongruent (ink color and the color word are different). The stimuli were randomized under the constraint that both the color word and the ink color did not repeat in two consecutive trials. Participants performed an interval-based version of the Stroop task (Figure 2A) in which they completed as many trials as they wished during a fixed time interval (8-12s). Previous work has used this task to demonstrate reliable adjustments in both drift rates and thresholds in response to incentives, and participant’s performance was well captured by a reward-rate model (Leng et al., 2021). Prior to each interval, participants saw a cue (1.5 s) instructing them to perform the task either as quickly (speed goal) or as accurately (accuracy goal) as possible (Forstmann et al., 2008; Ratcliff & Rouder, 1998). Participants received feedback (1.5 s) on the number of correct responses they made in the current interval. Crucially, there were two types of blocks (Figure 2A) which yielded different levels of control adjustment costs. For half of the blocks, the instructed condition was fixed over the entire block (fixed blocks), and in these blocks participants always received a cue telling them to focus on the same dimension of performance (e.g., speed). The other half of the blocks were varying blocks, in which performance goals changed within a block randomly. Before each block, participants were informed whether this will be a fixed or a varying block, these blocks were intermixed, and

their order was counterbalanced across participants. Participants performed 4 blocks, each of which included 20 intervals. Prior to the main task, participants performed practice blocks in which they first practiced the color-button mapping for the Stroop task, and then completed the task with the speed or accuracy condition and got familiarized with the cues for these conditions. We implemented the experiment in Psiturk (Gureckis et al., 2016) and participants performed the task on their own computers and were required to have a keyboard.

Study 2

The task was the same as in Study 1, but participants had 3 performance goals: speed, accuracy, and speed/accuracy. In the additional speed/accuracy goal, they were instructed to perform the task as fast and as accurately as possible. The experiment included 6 blocks, each of which included 20 intervals. Half of the blocks were fixed, meaning that participants performed under only one of the 3 goals in each of those blocks (1 block per goal). The other half of the blocks were varying. In these blocks participants switched between the three performance goals randomly.

Study 3

Participants performed the same Stroop task described in Study 1. They performed the task as fast (speed goal) or as accurately as possible (accuracy goal), and the task included varying and fixed blocks. Participants completed 6 blocks of this task, and each block consisted of 18 intervals. For both fixed and varying blocks, half of the blocks had a shorter time between the goal cue and task onset, and half of the blocks had longer time. Following the cue presentation (750ms), participants were presented with a fixation cross for either 250ms (short stimulus-onset asynchrony; SOA) or 1000ms (long SOA), immediately followed by task onset.

Study 4

Participants performed the same task as in Study 1, in 5 blocks consisting of 16 intervals. The task included 4 different block types. Participants performed 2 fixed goal blocks (1 block of each goal) in which the performance goal never changed. In varying blocks performance goals changed every 8, 4, or 2 intervals, and participants performed 1 block of each type. Following the cue presentation (1500 ms), participants were presented with a fixation cross for either 1250ms, immediately followed by task onset.

Statistical Analyses and Drift Diffusion Modeling

Behavioral Performance

Study 1. To predict reaction times and accuracies on each trial we fitted Bayesian multilevel regressions with goal type (speed vs. accuracy), block type (fixed vs. varying), and their interaction as predictors, while also controlling for the effects of congruency, interval length, and the time spent in the interval. For the reaction times analyses, we included only correct responses, and we did not analyze responses faster than 250ms and longer than 3000ms. All of the fixed predictors were allowed to vary across subjects. Regression models were fitted in R using the *brms* (Bürkner, 2017; Nalborczyk et al., 2019) package which relies on *Stan* (Carpenter et al., 2017) to implement Markov Chain Monte Carlo (MCMC) algorithm and estimate posterior distributions of parameters. We used ex-Gaussian likelihood to predict reaction times, and Bernoulli to predict accuracy, and the models included uniform priors. For each of the regression models we ran four MCMC simulations ("chains"; 20000 iterations per chain; 19000 warmup), and confirmed convergence by examining trace plots, autocorrelation, variance between and within chains (Gelman & Rubin, 1992), and posterior predictive checks. We summarized the posteriors for each parameter by reporting the mean of the posterior distribution (*b*) and the credible intervals (95% CrI). For inference we relied on the proportion of

the posterior samples on one side of 0 (e.g., $p_{b<0} = 0.01$) which represents the probability that given parameter estimate is below 0.

Study 2. Models and the model-fitting procedure was the same as in Study 1. The only difference was that the goal type now included 3 levels: speed, accuracy, and speed/accuracy.

Study 3. Models and the model-fitting procedure was the same as in Study 1. In addition to the predictors used in Study 1, here we also included the effect of SOA (250ms vs. 1000ms) as the main effect, as well as the interaction between goal type, block type, and SOA.

Study 4. Models and the model-fitting procedure was the same as in Study 1. Since this study included 4 different block types (fixed, switch every 8 intervals, switch every 4 intervals, switch every 2 intervals), the effect of block type was coded with an orthogonal polynomial contrast of degree 4. This allowed us to investigate the linear effect of the increase in switch frequency.

Drift diffusion Model

Study 1. We fitted hierarchical Bayesian drift diffusion models (Wiecki et al., 2013) to reaction time and accuracy data. For the error trials, we only included trials on which the incorrect response corresponded to the color word and thus driven by the automatic prepotent response to the word. This was done in order to exclude errors which are likely driven by random responding (cf., Leng et al., 2021). Fitted models included the effect of congruency on drift rate, as well as the effects of instruction type (speed vs. accuracy), block type (varying vs. fixed), and the interaction between instruction and block type on drift rate. We included the same effects on the threshold, but without the effect of congruency. We also included the effect of running time within interval on the threshold. Finally, we also included group and subject-level parameters of non-decision time, drift rate and non-decision time variability, and the probability of the response

coming from another distribution (random response). To ensure the reliability of our findings, we also fitted two additional models: one in which we allowed for non-decision times to vary across different performance goals and block types, and one in which we allowed the performance goal to influence the collapsing bound (Fengler et al., 2021). This analysis showed that the effects we report on drift rate and threshold do not qualitatively change in these more complex models (see Table S2). We ran 5 MCMC simulations (chains; 14000 iterations per chain; 12000 warmup) to estimate the model parameters. We confirmed chain convergence by examining trace plots, the ratio of variances between and within chains (Gelman-Rubin statistic), and posterior predictive checks. As for the regression models, we report the mean of the posterior distribution, the credible intervals, and the probability that the parameter of interest is below 0. All of the raw data and analysis scripts are available on GitHub (https://github.com/igrahek/ControlAdjustmentCosts.git).

Study 2. The models we used the model-fitting procedure was the same as in Study 1. Since there were 3 different goals in this study, the goal type predictor now included 3 levels: speed, accuracy, and speed/accuracy.

Study 3. We used the same models as in Study 1 and fitted them in the same way. We added the additional effect of SOA (250ms vs. 1000ms) as a main effect, as well as the interaction between goal type, block type, and SOA. This was done for both the drift rates and thresholds.

Study 4. We used the same model as in Study 1. Since this study included 4 different block types (fixed, switch every 8 intervals, switch every 4 intervals, switch every 2 intervals), the effect of block type was included as a linear, quadratic, and a cubic effect.

Computational Model

We used the computational model described in the "Control Adjustment Costs: A Dynamical Systems Model" section to simulate performance of an artificial agent in each of the 4 studies. For each of the studies we had our agent perform the task which was closely matched to the task performed by the participants. The agent performed the same number of blocks, block types, performance goals within a block, and the number of intervals in the block as our participants. We report in the results section the average drift rates and thresholds that the model generated for each of the experimental conditions. To simulate task performance, the agent first found the optimal control state (Eq. 1) and then gradually moved toward that state (Eq. 2).

Determining Optimal Control Intensity

We rely on a reward-rate optimization model (Bogacz et al., 2006; Leng et al., 2021) to specify how optimal control states for each of the performance goals across our experiments are determined. This model (Eq. 3) calculates the expected reward rate (RR) given the current value of giving a correct response (w_1) and making an error (w_2). These weights are specified by the current performance goals (e.g., valuing speed over accuracy) or incentives in the environment (e.g., rewards for correct responses). The weights scale the rate of correct responses ($1 - ER$) and the error rate (ER) obtained from the Drift Diffusion model (Bogacz et al., 2006). The expected reward rate includes two types of costs. The opportunity cost (Kurzban et al., 2013; Otto & Daw, 2019) which discounts the overall outcome and includes both the decision time (DT) and non-decision time (t). The second cost is the intensity cost (Shenhav et al., 2013) which represents the quadratic cost on cognitive control intensity, operationalized as the drift rate (v). The expected reward rate is thus calculated as:

$$RR(v, a, w_1, w_2, t) = \frac{w_1 \times (1 - ER(v,a)) - w_2 \times ER(v,a)}{(DT(v,a) + t)} - v^2 \quad (3)$$

The model finds optimal levels of drift rate (v^*) and threshold (a^*) which maximize the expected reward rate:

$$v^*, a^* = argmax_{v,a}[RR|w_1, w_2, t] \quad (4)$$

To match our simulations to the range of participant's performance, we determined the weights for the value of giving a correct response ($\boldsymbol{w_1}$) and the value of committing an error ($\boldsymbol{w_2}$) based on the empirical estimates of drift rates, thresholds, and non-decision times in fixed blocks. We employed an inverse reward-rate optimization procedure (cf. Leng et al., 2021) which relies on the reward-rate optimal model (Eq. 1) to find the weights which produce the empirically observed control signals. In this way we determined the two weights for each of the performance goals in each of the studies. On each interval, optimal levels of drift rate (v^*) and threshold (a^*) were found by maximizing the expected reward rate given the weights for the current performance goal and the non-decision time (Eq. 1 & 2).

Control Adjustment Dynamics

On each interval, after the presentation of the performance cue which determined which weights should be used (e.g., speed vs. accuracy), drift rates and thresholds were gradually adjusted from their starting levels (v_0, a_0), toward the reward-optimal levels (v^*, a^*) described above. The initial values of drift rate and threshold ($v = 3;\ a = 0.8$), the speed of adjustment ($\tau = 0.2;\ dt = 0.005$), as well as the noise level ($c = 0.1$) were fixed across all simulations. Due to the use of the sigmoid activation function, drift rates and thresholds were normalized to the 0-1 range before they were adjusted, and they were re-normalized to the original range after the adjustments. The adjustments were performed in discrete steps, and the number of steps was determined by the SOA between cue offset and task onset in each study. At each step, the reward-

optimal levels of drift rate and threshold (v^*, a^*) were passed through a sigmoid activation function with a fixed gain ($g = 5$) and bias ($\theta = 0.5$).

Parameters for Each Study

Study 1. We simulated behavior on intervals with the speed goal ($w_1 = 8.62, w_2 = 30.39$) and the accuracy goal ($w_1 = 12.17, w_2 = 916.71$). The non-decision time was set to the estimate in the fixed blocks obtained from the drift diffusion model fit ($t = 0.32$). The number of steps for control adjustment was 1500, matched to the SOA in Study 1 (1500 ms). We simulated 4 blocks which included 30 intervals with 10 trials in each interval. Half of the blocks were fixed, meaning that the performance goal, and the corresponding weights, never changed. The other half of blocks included performance goal changes randomly between intervals.

Study 2. Behavior was simulated on intervals with the speed goal ($w_1 = 11.90, w_2 = 32.98$), the accuracy goal ($w_1 = 17.88, w_2 = 3673.4$), and the speed & goal ($w_1 = 15.19, w_2 = 435.97$). The non-decision time was set to ($t = 0.33$). The number of steps (1500) was matched to the SOA in Study 2 (1500 ms). We simulated 6 blocks which included 20 intervals with 10 trials in each interval. Half of the blocks were fixed, meaning that the performance goal, and the corresponding weights, never changed. The other half of blocks included performance goal changes randomly between intervals.

Study 3. Simulations included the speed goal ($w_1 = 15.65, w_2 = 145.76$) and the accuracy goal ($w_1 = 16.58, w_2 = 671.97$). The non-decision time was set to$t = 0.39$. The number of steps was matched to the time between the cue onset and the first trial onset in Study 3 (short SOA = 1000 ms; long SOA = 1750 ms). We simulated 6 blocks which included 18 intervals with 10 trials in each interval. Half of the blocks were fixed, meaning that the

performance goal, and the corresponding weights, never changed. The other half of blocks included changes between the two performance goals every 1 interval.

Study 4. This simulation included two performance goals: the speed goal ($w_1 = 17.79, w_2 = 136.07$) and the accuracy goal ($w_1 = 20.50, w_2 = 1787.4$). The non-decision time was set to: $t = 0.38$. The number of steps for control adjustment was 2750 matched the time from cue onset until the first trial onset in Study 4 (2750 ms). We simulated 5 blocks which included 16 intervals with 10 trials in each interval. The blocks were divided into 4 groups: fixed blocks (2 blocks; always the same performance goal), switch every 8 intervals (1 block), switch every 4 intervals (1 block), and switch every 2 intervals (1 block). The gain parameters differed across the fixed blocks ($g = 5$), and the switch 8 ($g = 3$), switch 4 ($g = 2.5$), and switch 2 blocks ($g = 2$).

Acknowledgements

This work was supported by the Training Program for Interactionist Cognitive Neuroscience T32-MH115895 (X.L.), NIMH R01MH124849 (A.S.), NIMH R21MH122863 (A.S.), NSF Career Award 204611 (A.S.), and NIH Award 1S10OD02518. The authors thank Jonathan D. Cohen, Alexander Fengler, Atsushi Kikumoto, Harrison Ritz, and all members of the Shenhav lab for many fruitful discussions of this work. We thank Mahalia Prater Fahey and Debbie Yee for providing the datasets on which some of the ideas in this manuscript were initially tested. We thank Ainsley Bonin for her contributions to the project during her summer internship at the Shenhav lab.

Supplementary materials

Table S1

Study 1 parameter estimates for hierarchical regressions predicting behavioral performance (reaction times and accuracy) and parameter estimates of the Drift Diffusion Model (a – threshold; v – drift rate; t – non-decision time; z – bias).

Parameter	***Estimate***	***95% Credible Interval***	***Posterior probability (p<0)***
Reaction times (ms)			
Intercept	718.62	682.05, 755.69	0.00
Congruency (I-C)	39.18	28.82, 49.61	0.00
Goal (A-S)	128.25	85.36, 172.52	0.00
Block (V-F)	-12.09	-27.66, 4.04	0.93
Goal * Block	28.09	3.54, 51.41	0.01
Interval number	-7.72	-11.48, -3.92	1.00
Interval length	3.81	2.01, 5.58	0.00
Time in interval	-22.71	-24.68, -20.74	1.00
Accuracy (log odds)			
Intercept	1.66	1.35, 1.99	0.00
Congruency (I-C)	-0.25	-0.36, -0.14	1.00
Goal (A-S)	1.67	1.12, 2.22	0.00
Block (V-F)	-0.11	-0.29, 0.07	0.89
Goal * Block	0.44	0.12, 0.76	0.00
Interval number	-0.17	-0.23, -0.12	1.00
Interval length	-0.03	-0.05, 0.00	0.96
Time in interval	-0.01	-0.04, 0.02	0.80
Drift Diffusion Model Estimates			
(a) Intercept	0.94	0.87, 1.01	0.00
(a) Goal (A-S)	0.42	0.27, 0.58	0.00
(a) Block (V-F)	0.00	-0.08, 0.08	0.49
(a) Goal * Block	0.08	0.03, 0.13	0.00
(a) Time in interval	-0.07	-0.08, -0.06	0.00
(v) Intercept	2.66	2.49, 2.84	0.00
(v) Goal (A-S)	0.54	0.31, 0.77	0.00
(v) Block (V-F)	0.03	-0.16, 0.09	0.29
(v) Goal * Block	0.05	-0.12, 0.21	0.28
(v) Congruency (I-C)	-0.34	-0.41, -0.28	0.00
(t) Intercept	0.32	0.3, 0.34	0.00
(z) Congruency (I-C)	-0.02	-0.03, -0.01	0.00
Lapse rate	0.00	0.00, 0.00	0.00

Table S2

Study 1 parameter estimates for parameter estimates of the Drift Diffusion Models which include:1) effects on non-decision time, and 2) a collapsing bound (a – threshold; v – drift rate; t – non-decision time; z – bias).

Parameter	***Estimate***	***95% Credible Interval***	***Posterior probability (p<0)***
Drift Diffusion Model Estimates with Effects on non-decision time			
(a) Intercept	0.86	0.80, 0.93	0.00
(a) Goal (S-A)	0.30	0.18, 0.41	0.00
(a) Block (V-F)	0.00	-0.06, 0.06	0.44
(a) Goal * Block	0.05	0.01, 0.10	0.01
(a) Time in interval	-0.07	-0.08, -0.06	0.00
(v) Intercept	2.38	2.20, 2.56	0.00
(v) Goal (A-S)	0.51	0.30, 0.73	0.00
(v) Block (V-F)	-0.03	-0.14, 0.10	0.34
(v) Goal * Block	0.04	-0.12, 0.20	0.29
(v) Congruency (I-C)	-0.31	-0.37, -0.25	0.00
(t) Intercept	0.34	0.32, 0.37	0.00
(t) Goal (A-S)	0.04	0.02, 0.07	0.00
(t) Block (V-F)	0.00	-0.02, 0.02	0.41
(t) Goal * Block	0.01	0.00, 0.03	0.10
(z) Congruency (I-C)	-0.03	-0.04, 0.02	0.00
Lapse rate	0.00	0.00, 0.00	0.00
Drift Diffusion Model Estimates with a Collapsing Bound			
(a) Intercept	0.55	0.50, 0.60	0.00
(a) Goal (A-S)	0.20	0.12, 0.28	0.00
(a) Block (V-F)	0.00	-0.04, 0.05	0.46
(a) Goal * Block	0.04	0.01, 0.07	0.00
(v) Intercept	1.87	1.72, 2.02	0.00
(v) Goal (A-S)	0.50	0.31, 0.68	0.00
(v) Block (V-F)	-0.02	-0.11, 0.08	0.36
(v) Goal * Block	0.05	-0.07, 0.17	0.20
(v) Congruency (I-C)	-0.34	-0.39, -0.30	0.00
(t) Intercept	0.26	0.23, 0.28	0.00
(t) Goal (S-A)	-0.02	-0.03, -0.01	0.00
(theta) Intercept	0.16	-0.08, 0.34	0.12
(theta) Goal (S-A)	-0.10	-0.19, -0.01	0.01
Lapse rate	0.01	0.01, 0.01	0.00

Table S3

Study 1 – switch vs. repeat intervals analysis. Parameter estimates for hierarchical regressions predicting behavioral performance (reaction times and accuracy) and parameter estimates of the Drift Diffusion Model (a – threshold; v – drift rate; t – non-decision time; z – bias).

Parameter	***Estimate***	***95% Credible Interval***	***Posterior probability (p<0)***
Reaction times (ms)			
Intercept	724.11	685.74, 759.48	0.00
Congruency (I-C)	40.27	28.69, 51.85	0.00
Goal (A-S)	113.19	71.51, 156.99	0.00
Switch (S-R)	-4.31	-10.81, 2.19	0.91
Goal * Switch	10.97	-1.65, 23.86	0.04
Interval number	-7.25	-12.46, -2.15	1.00
Interval length	1.54	-1.04, 4.18	0.12
Time in interval	-22.67	-25.52, -19.83	1.00
Accuracy (log odds)			
Intercept	1.77	1.43, 2.11	0.00
Congruency (I-C)	-0.28	-0.42, -0.14	1.00
Goal (A-S)	1.45	0.87, 2.02	0.00
Switch (S-R)	0.08	-0.03, 0.19	0.07
Goal * Switch	0.23	0.01, 0.46	0.02
Interval number	-0.19	-0.27, -0.11	1.00
Interval length	-0.05	-0.10, -0.01	0.99
Time in interval	-0.01	-0.05, 0.03	0.73
Drift Diffusion Model Estimates			
(a) Intercept	0.94	0.86, 1.01	0.00
(a) Goal (A-S)	0.37	0.23, 0.51	0.00
(a) Block (S-R)	0.00	-0.04, 0.04	0.45
(a) Goal * Block	0.04	0.00, 0.07	0.02
(a) Time in interval	-0.14	-0.16, -0.12	0.00
(v) Intercept	2.71	2.49, 2.94	0.00
(v) Goal (A-S)	0.46	0.23, 0.69	0.00
(v) Block (S-R)	-0.01	-0.12, 0.09	0.42
(v) Goal * Block	0.00	-0.15, 0.15	0.49
(v) Congruency (I-C)	-0.34	-0.42, -0.27	0.00
(t) Intercept	0.32	0.30, 0.35	0.00
(z) Congruency (I-C)	-0.02	-0.03, -0.01	0.00
Lapse rate	0.00	0.00, 0.00	0.00

Table S4

Study 2 parameter estimates for hierarchical regressions predicting behavioral performance (reaction times and accuracy) and parameter estimates of the Drift Diffusion Model (a – threshold; v – drift rate; t – non-decision time; z – bias).

Parameter	***Estimate***	***95% Credible Interval***	***Posterior probability (p<0)***
Reaction times (ms)			
Intercept	669.25	641.01, 697.21	0.00
Congruency (I-C)	39.30	31.53, 47.01	0.00
Goal (A-S)	143.32	95.18, 190.48	0.00
Goal (A-S/A)	-52.74	-79.10, -25.10	1.00
Block (V-F)	-3.17	-17.84, 11.35	0.67
Goal (A-S) * Block	30.30	-3.83, 64.80	0.04
Goal (A-S/A) * Block	-8.43	-35.40, 17.49	0.74
Interval number	-9.20	-12.91, -5.50	1.00
Interval length	1.91	0.18, 3.65	0.02
Time in interval	-21.74	-23.71, -19.81	1.00
Accuracy (log odds)			
Intercept	1.83	1.45, 2.21	0.00
Congruency (I-C)	-0.32	-0.50, -0.14	1.00
Goal (S-A)	2.01	1.50, 2.52	0.00
Goal (A-S/A)	-0.69	-0.94, -0.45	1.00
Block (V-F)	-0.06	-0.23, 0.12	0.75
Goal (A-S) * Block	0.75	0.31, 1.20	0.00
Goal (A-S/A) * Block	-0.27	-0.64, 0.10	0.92
Interval number	0.01	-0.04, 0.07	0.33
Interval length	-0.02	-0.05, 0.01	0.92
Time in interval	-0.06	-0.09, -0.03	1.00
Drift Diffusion Model Estimates			
(a) Intercept	0.89	0.82, 0.97	0.00
(a) Goal (A-S)	0.48	0.38, 0.58	0.00
(a) Goal (A-S/A)	0.20	0.10, 0.31	0.00
(a) Block (V-F)	0.00	-0.06, 0.06	0.51
(a) Goal (A-S) * Block	0.11	0.04, 0.18	0.00
(a) Goal (A-S/A) * Block	0.04	-0.03, 0.11	0.17
(a) Time in interval	-0.07	-0.08, -0.05	0.00
(v) Intercept	3.20	2.91, 3.49	0.00
(v) Goal (A-S)	0.66	0.47, 0.86	0.00
(v) Goal (A-S/A)	0.27	0.09, 0.46	0.01
(v) Block (V-F)	0.07	-0.06, 0.20	0.18
(v) Goal (A-S) * Block	0.36	0.16, 0.55	0.00
(v) Goal (A-S/A) * Block	0.09	-0.09, 0.27	0.21
(v) Congruency (I-C)	-0.41	-0.50, -0.33	0.00
(t) Intercept	0.33	0.31, 0.36	0.00
(z) Congruency (I-C)	-0.02	-0.03, -0.02	0.00
Lapse rate	0.00	0.00, 0.00	0.00

Table S5

Study 3 parameter estimates for hierarchical regressions predicting behavioral performance (reaction times and accuracy) and parameter estimates of the Drift Diffusion Model (a – threshold; v – drift rate; t – non-decision time; z – bias).

Parameter	***Estimate***	***95% Credible Interval***	***Posterior probability (p<0)***
Reaction times (ms)			
Intercept	746.21	706.34, 785.06	0.00
Congruency (I-C)	43.8	37.32, 50.49	0.00
SOA (1000-250)	2.35	-8.05, 12.57	0.32
Goal (A-S)	50.66	29.08, 71.58	0.00
Block (V-F)	-2.09	-14.41, 9.56	0.63
SOA * Goal	-12.65	-24.03, -1.33	0.99
SOA * Block	-2.24	-17.46, 13.09	0.62
Goal * Block	14.37	-1.68, 29.37	0.04
SOA * Goal * Block	-19.55	-51.10, 11.65	0.90
Interval number	-7.83	-12.43, -3.07	1.00
Interval length	4.18	2.48, 5.90	0.00
Time in interval	-33.45	-35.38, -31.58	1.00
Accuracy (log odds)			
Intercept	2.32	2.02, 2.62	0.00
Congruency (I-C)	-0.37	-0.49, -0.26	1.00
SOA (1000-250)	0.04	-0.14, 0.22	0.31
Goal (A-S)	0.88	0.57, 1.20	0.00
Block (V-F)	0.13	-0.04, 0.31	0.07
SOA * Goal	-0.02	-0.24, 0.19	0.59
SOA * Block	0.01	-0.24, 0.26	0.46
Goal * Block	0.62	0.36, 0.87	0.00
SOA * Goal * Block	-0.20	-0.61, 0.19	0.84
Interval number	-0.08	-0.16, -0.01	0.98
Interval length	-0.01	-0.04, 0.03	0.69
Time in interval	-0.05	-0.09, -0.02	1.00
Drift Diffusion Model Estimates			
(a) Intercept	0.91	0.83, 1.00	0.00
(a) SOA (1000-250)	-0.01	-0.03, 0.02	0.28
(a) Goal (A-S)	0.18	0.12, 0.24	0.00
(a) Block (V-F)	0	-0.03, 0.03	0.42
(a) Goal (A-S) * SOA	-0.03	-0.07, 0.01	0.06
(a) Block * SOA	-0.02	-0.06, 0.03	0.24
(a) Goal (A-S) * Block	0.07	0.03, 0.11	0.00
(a) Goal (A-S) * Block * SOA	0.07	0.00, 0.15	0.03
(a) Time in interval	-0.10	-0.12, -0.09	0.00
(v) Intercept	3.02	2.74, 3.29	0.00
(v) SOA (1000-250)	-0.08	-0.19, 0.03	0.07
(v) Goal (A-S)	0.14	0.06, 0.23	0.00
(v) Block (V-F)	-0.07	-0.17, 0.02	0.06
(v) Goal (A-S) * SOA	0.02	-0.13, 0.17	0.41
(v) Block * SOA	-0.03	-0.18, 0.11	0.32
(v) Goal (A-S) * Block	0.03	-0.09, 0.16	0.34
(v) Goal (A-S) * Block * SOA	-0.09	-0.36, 0.17	0.26
(v) Congruency (I-C)	-0.34	-0.42, -0.26	0.00
(t) Intercept	0.39	0.37, 0.41	0.00
(z) Congruency (I-C)	-0.03	-0.04, -0.02	0.00

Lapse rate	0.00	0.00, 0.00	0.00

Table S6

Study 4 parameter estimates for hierarchical regressions predicting behavioral performance (reaction times and accuracy) and parameter estimates of the Drift Diffusion Model (a – threshold; v – drift rate; t – non-decision time; z – bias).

Parameter	***Estimate***	***95% Credible Interval***	***Posterior probability (p<0)***
Reaction times (ms)			
Intercept	728.8	683.41, 772.02	0.00
Congruency (I-C)	40.04	32.34, 47.86	0.00
Goal (A-S)	54.14	37.96, 70.30	0.00
Switch frequency (linear effect)	-3.01	-16.15, 9.71	0.67
Switch frequency (quadratic effect)	-6.73	-17.44, 4.39	0.90
Switch frequency (cubic effect)	4.39	-3.51, 11.95	0.12
Switch frequency (linear effect) *Goal	25.20	8.50, 42.58	0.00
Switch frequency (quadratic effect) *Goal	20.98	8.22, 33.39	0.00
Switch frequency (cubic effect) *Goal	3.03	-8.45, 14.90	0.31
Interval number	-12.89	-17.33, -8.62	1.00
Interval length	2.94	1.72, 4.18	0.00
Time in interval	-13.76	-15.03, -12.47	1.00
Accuracy (log odds)			
Intercept	2.49	2.25, 2.70	0.00
Congruency (I-C)	-0.25	-0.35, -0.15	1.00
Goal (A-S)	0.80	0.58, 1.04	0.00
Switch frequency (linear effect)	-0.24	-0.39, -0.10	1.00
Switch frequency (quadratic effect)	-0.20	-0.35, -0.07	1.00
Switch frequency (cubic effect)	0.01	-0.06, 0.09	0.35
Switch frequency (linear effect) * Goal	0.06	0.00, 0.12	0.03
Switch frequency (quadratic effect) * Goal	-0.03	-0.06, 0.00	0.96
Switch frequency (cubic effect) * Goal	-0.09	-0.12, -0.07	1.00
Interval number	0.48	0.25, 0.74	0.00
Interval length	0.21	0.00, 0.42	0.03
Time in interval	0.09	-0.09, 0.27	0.15
Drift Diffusion Model Estimates			
(a) Intercept	0.96	0.89, 1.03	0.00
(a) Goal (A-S)	0.23	0.17, 0.30	0.00
(a) Switch frequency (linear effect)	-0.05	-0.08, -0.02	0.00
(a) Switch frequency (quadratic effect)	-0.06	-0.09, -0.03	0.00
(a) Switch frequency (cubic effect)	0.00	-0.03, 0.02	0.47
(a) Switch frequency (linear effect) * Goal	-0.12	-0.21, -0.04	0.00
(a) Switch frequency (quadratic effect) * Goal	-0.02	-0.04, 0.00	0.06
(a) Switch frequency (cubic effect) * Goal	0.00	-0.02, 0.02	0.47
(a) Time in interval	-0.05	-0.06, -0.04	0.00
(v) Intercept	3.38	3.18, 3.57	0.00
(v) Goal (A-S)	0.30	0.17, 0.44	0.00
(v) Switch frequency (linear effect)	0.01	-0.08, 0.10	0.41
(v) Switch frequency (quadratic effect)	-0.08	-0.16, 0.00	0.02
(v) Switch frequency (cubic effect)	-0.01	-0.08, 0.05	0.34
(v) Switch frequency (linear effect) * Goal	-0.20	-0.43, 0.05	0.05
(v) Switch frequency (quadratic effect) * Goal	0.01	-0.05, 0.07	0.39
(v) Switch frequency (cubic effect) * Goal	0.05	-0.01, 0.11	0.06
(v) Congruency (I-C)	-0.38	-0.45, -0.32	0.00
(t) Intercept	0.38	0.36, 0.41	0.00
(z) Congruency (I-C)	-0.02	-0.03, -0.01	0.00

Lapse rate	0.00	0.00, 0.00	0.00

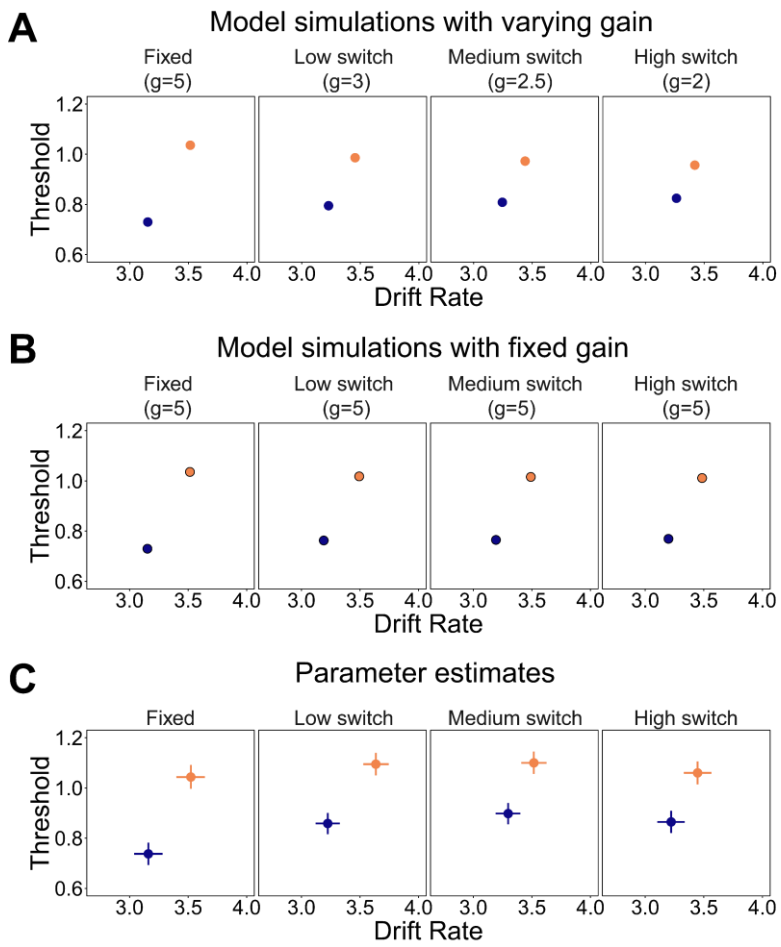

Figure S1. Comparison between models with and without varying gain for Study 4. A. Simulations from the model which assumes that participants proactively adjust gain based on the instructions informing them that within a given block goals will be fixed, or will change every 8 (low), 4 (medium), or 2 (high) intervals. This model captures the empirical effect well. **B. Simulations from the model which assumes that gain is fixed.** Model simulations across four levels of goal switch frequency with fixed gain. This model also predicts increasing control adjustment costs with the increase of the goal switch frequency, but the magnitude of this effect is substantially smaller than the empirically observed effect. **C.** Drift diffusion parameter estimates. Error bars represent 95% credible intervals.

References

Alexander, W. H., Brown, J. W., Collins, A. G. E., Hayden, B. Y., & Vassena, E. (2018). Prefrontal Cortex in Control: Broadening the Scope to Identify Mechanisms. *Journal of Cognitive Neuroscience*, *30*(8), 1061–1065. https://doi.org/10.1162/jocn_a_01154

Alport, A., Styles, E. A., & Shulan, H. (1994). Shifting Intentional Set: Exploring the Dynamic Control of Tasks. In *Attention and performance XV: Conscious and nonconscious information processing*. The MIT Press. https://doi.org/10.7551/mitpress/1478.001.0001

Altmann, E. M. (2007). Cue-independent task-specific representations in task switching: Evidence from backward inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 892–899. https://doi.org/10.1037/0278-7393.33.5.892

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, *120*(3), 338–375. https://doi.org/10.1037/0033-2909.120.3.338

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Botvinick, M., & Braver, T. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism. *Annual Review of Psychology*, *66*(1), 83–113. https://doi.org/10.1146/annurev-psych-010814-015044

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

Cools, R. (2015). The cost of dopamine for dynamic cognitive control. *Current Opinion in Behavioral Sciences*, *4*, 152–159. https://doi.org/10.1016/j.cobeha.2015.05.007

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Fengler, A., Govindarajan, L. N., Chen, T., & Frank, M. J. (2021). Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience. *ELife*, *10*, e65074. https://doi.org/10.7554/eLife.65074

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Cramon, D. Y. von, Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542. https://doi.org/10.1073/pnas.0805903105

Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21315-z

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472.

Gilbert, S. J., & Shallice, T. (2002). Task Switching: A PDP Model. *Cognitive Psychology*, *44*(3), 297–337. https://doi.org/10.1006/cogp.2001.0770

Golwitzer (Ed.). (2012). Mindset Theory of Action Phases. In *Midset Theory of Action Phases*. SAGE.

Goschke, T. (2013). Volition in Action: Intentions, Control Dilemmas, and the Dynamic Regulation of Cognitive Control. In *Volition in Action: Intentions, Control Dilemmas, and the Dynamic Regulation of Cognitive Control*. The MIT Press. https://doi.org/10.7551/mitpress/9780262018555.001.0001

Grahek, I., Frömer, R., Prater Fahey, M., & Shenhav, A. (2023). Learning when effort matters: Neural dynamics underlying updating and adaptation to changes in performance efficacy. *Cerebral Cortex*, *33*(5), 2395–2411. https://doi.org/10.1093/cercor/bhac215

Grahek, I., Musslick, S., & Shenhav, A. (2020). A computational perspective on the roles of affect in cognitive control. *International Journal of Psychophysiology*, *151*, 25–34. https://doi.org/10.1016/j.ijpsycho.2020.02.001

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. https://doi.org/10.3758/s13428-015-0642-8

Haggard, P. (2019). The Neurocognitive Bases of Human Volition. *Annual Review of Psychology*, *70*(1), 9–28. https://doi.org/10.1146/annurev-psych-010418-103348

Holroyd, C. (2023). *The Controllosphere: The Neural Origin of Cognitive Effort*. PsyArXiv. https://doi.org/10.31234/osf.io/2qmhg

Jersild, A. T. (1927). *Mental set and shift*. Columbia University.

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review*, *30*(7), 865–878. https://doi.org/10.1016/j.cpr.2010.03.001

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, *136*(5), 849–874. https://doi.org/10.1037/a0019842

Kikumoto, A., & Mayr, U. (2020). Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences*, *117*(19), 10603–10608. https://doi.org/10.1073/pnas.1922166117

Kim, J. Z., Soffer, J. M., Kahn, A. E., Vettel, J. M., Pasqualetti, F., & Bassett, D. S. (2018). Role of graph architecture in controlling dynamical networks with applications to neural systems. *Nature Physics*, *14*(1), Article 1. https://doi.org/10.1038/nphys4268

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking—An integrative review of dual-task and task-switching research. *Psychological Bulletin*, *144*(6), 557–583. https://doi.org/10.1037/bul0000144

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665–682. https://doi.org/10.1037/a0020198

Krebs, R. M., Boehler, C. N., Roberts, K. C., Song, A. W., & Woldorff, M. G. (2012). The Involvement of the Dopaminergic Midbrain and Cortico-Striatal-Thalamic Circuits in the Integration of Reward Prospect and Attentional Task Demands. *Cerebral Cortex*, *22*(3), 607–615. https://doi.org/10.1093/cercor/bhr134

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679. https://doi.org/10.1017/S0140525X12003196

Leng, X., Yee, D., Ritz, H., & Shenhav, A. (2021). Dissociable influences of reward and punishment on adaptive cognitive control. *PLOS Computational Biology*, *17*(12), e1009737. https://doi.org/10.1371/journal.pcbi.1009737

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLOS Computational Biology*, *14*(4), e1006043. https://doi.org/10.1371/journal.pcbi.1006043

Liljenström, H. (2003). Neural Stability and Flexibility: A Computational Approach. *Neuropsychopharmacology*, *28*(1), Article 1. https://doi.org/10.1038/sj.npp.1300137

Liu, C., & Yeung, N. (2020). Dissociating expectancy-based and experience-based control in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(2), 131–154. https://doi.org/10.1037/xhp0000704

Logan, G. D., & Bundesen, C. (2003). Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? *Journal of Experimental Psychology: Human Perception and Performance*, *29*(3), 575–599. https://doi.org/10.1037/0096-1523.29.3.575

Manohar, S. G., Chong, T. T.-J., Apps, M. A. J., Batla, A., Stamelou, M., Jarman, P. R., Bhatia, K. P., & Husain, M. (2015). Reward Pays the Cost of Noise Reduction in Motor and Cognitive Control. *Current Biology*, *25*(13), 1707–1716. https://doi.org/10.1016/j.cub.2015.05.038

Mayr, U., & Keele, S. W. (2000). Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General*, *129*(1), 4–26. https://doi.org/10.1037/0096-3445.129.1.4

Mayr, U., & Kliegl, R. (2000). Task-set switching and long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1124–1140. https://doi.org/10.1037/0278-7393.26.5.1124

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1423–1442. https://doi.org/10.1037/0278-7393.22.6.1423

Millan, M. J., Agid, Y., Brüne, M., Bullmore, E. T., Carter, C. S., Clayton, N. S., Connor, R., Davis, S., Deakin, B., DeRubeis, R. J., & others. (2012). Cognitive dysfunction in psychiatric disorders: Characteristics, causes and the quest for improved therapy. *Nature Reviews Drug Discovery*, *11*(2), 141–168.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Musslick, S., Bizyaeva, A., Agaron, S., Leonard, N., & Cohen, J. D. (2019). Stability-Flexibility Dilemma in Cognitive Control: A Dynamical System Perspective. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. https://par.nsf.gov/biblio/10125021-stability-flexibility-dilemma-cognitive-control-dynamical-system-perspective

Musslick, S., Botvinick, M. M., Shenhav, A., & Cohen, J. D. (2015). *A computational model of control allocation based on the Expected Value of Control*. 6.

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775. https://doi.org/10.1016/j.tics.2021.06.001

Musslick, S., Jang, S. J., Shvartsman, M., Shenhav, A., & Cohen, J. D. (2018). *Constraints associated with cognitive control and the stability-flexibility dilemma*. 6.

Musslick, S., & Masís, J. (2023). Pushing the Bounds of Bounded Optimality and Rationality. *Cognitive Science*, *47*(4), e13259. https://doi.org/10.1111/cogs.13259

Nalborczyk, L., Batailler, C., L, œvenbruck H., Vilain, A., & B, ürkner P.-C. (2019). An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230. https://doi.org/10.1016/j.jmp.2009.02.003

O'Reilly, R. C. (2020). Unraveling the Mysteries of Motivation. *Trends in Cognitive Sciences*, *24*(6), 425–434. https://doi.org/10.1016/j.tics.2020.03.001

Otto, A. R., & Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia*, *123*, 92–105. https://doi.org/10.1016/j.neuropsychologia.2018.05.006

Padmala, S., & Pessoa, L. (2011). Reward Reduces Conflict by Enhancing Attentional Control and Biasing Visual Cortical Processing. *Journal of Cognitive Neuroscience*, *23*(11), 3419–3432. https://doi.org/10.1162/jocn_a_00011

Parro, C., Dixon, M. L., & Christoff, K. (2018). The neural basis of motivational influences on cognitive control. *Human Brain Mapping*, *39*(12), 5097–5111. https://doi.org/10.1002/hbm.24348

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Ritz, H., Leng, X., & Shenhav, A. (2022). Cognitive Control as a Multivariate Optimization Problem. *Journal of Cognitive Neuroscience*, *34*(4), 569–591. https://doi.org/10.1162/jocn_a_01822

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A Network Model of Catecholamine Effects: Gain, Signal-to-Noise Ratio, and Behavior. *Science*, *249*(4971), 892–895. https://doi.org/10.1126/science.2392679

Shallice, T., & Burgess, P. W. (1991). Deficits in Strategy Application Following Frontal Love Damage in Man. *Brain*, *114*(2), 727–741. https://doi.org/10.1093/brain/114.2.727

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, *79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, *40*(1), 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Simen, P., Cohen, J. D., & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks*, *19*(8), 1013–1026. https://doi.org/10.1016/j.neunet.2006.05.038

Steyvers, M., Hawkins, G. E., Karayanidis, F., & Brown, S. D. (2019). A large-scale analysis of task switching practice effects across the lifespan. *Proceedings of the National Academy of Sciences*, *116*(36), 17735–17740. https://doi.org/10.1073/pnas.1906788116

Ueltzhöffer, K., Armbruster-Genç, D. J. N., & Fiebach, C. J. (2015). Stochastic Dynamics Underlying Cognitive Stability and Flexibility. *PLOS Computational Biology*, *11*(6), e1004331. https://doi.org/10.1371/journal.pcbi.1004331

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Verguts, T., Vassena, E., & Silvetti, M. (2015). Adaptive effort investment in cognitive and physical tasks: A neurocomputational model. *Frontiers in Behavioral Neuroscience*, *9*, 57. https://doi.org/10.3389/fnbeh.2015.00057

Wasylyshyn, C., Verhaeghen, P., & Sliwinski, M. J. (2011). Aging and task switching: A meta-analysis. *Psychology and Aging*, *26*(1), 15–20. https://doi.org/10.1037/a0020912

Waszak, F., Hommel, B., & Allport, A. (2004). Semantic generalization of stimulus-task bindings. *Psychonomic Bulletin & Review*, *11*(6), 1027–1033. https://doi.org/10.3758/BF03196732

Weninger, L., Srivastava, P., Zhou, D., Kim, J. Z., Cornblath, E. J., Bertolero, M. A., Habel, U., Merhof, D., & Bassett, D. S. (2022). Information content of brain states is explained by structural constraints on state energetics. *Physical Review E*, *106*(1), 014401. https://doi.org/10.1103/PhysRevE.106.014401

Westbrook, J. A., Bosch, R. van den, Määttä, J. I., Hofmans, L., Papadopetraki, D., Cools, R., & Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus

costs of cognitive work. *Science*, *367*(6484), 1362–1366. https://doi.org/10.1126/science.aaz5891

Westbrook, J. A., Kester, D., & Braver, T. S. (2013). What Is the Subjective Cost of Cognitive Effort? Load, Trait, and Aging Effects Revealed by Economic Preference. *PLOS ONE*, *8*(7), e68210. https://doi.org/10.1371/journal.pone.0068210

Wiecki, T., Sofer, I., & Frank, M. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. https://www.frontiersin.org/article/10.3389/fninf.2013.00014

Wylie, G., & Allport, A. (2000). Task switching and the measurement of "switch costs." *Psychological Research*, *63*(3), 212–233. https://doi.org/10.1007/s004269900003