Adaptive decision policy dynamics

A DISSERTATION PRESENTED
BY
Krista Bond

COMMITTEE
Timothy D. Verstynen, Chair
Lori L. Holt
Michael J. Tarr

IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
COGNITIVE NEUROSCIENCE

DEPARTMENT OF PSYCHOLOGY
CENTER FOR THE NEURAL BASIS OF COGNITION
CARNEGIE MELLON NEUROSCIENCE INSTITUTE
CARNEGIE MELLON UNIVERSITY

MAY, 2022

Abstract

The world changes. Therefore, successful adaptation requires flexible decision-making. Knowing when to rely on what is known (exploit) or sample alternatives (explore) is crucial to this flexibility, as this decision influences the knowledge we have about the world. Previous work on adaptive exploration has largely been studied at the level of action-value estimation, with the focus on the actions themselves[68,122,61]. This approach has been fruitful, allowing us to begin to taxonomize the kinds of exploration that exist and the contexts under which they operate, and this has helped us to take first steps toward understanding the influence of neuromodulatory systems on these taxonomized forms of exploration[64,161,177,162]. However, these results concentrate on behavioral outcomes rather than how a learner should change the set of rules that shape whether one should explore new options or exploit what they know (i.e. decision policy). So, investigating how underlying decision dynamics shift under changing conditions to adaptively navigate the exploration-exploitation continuum is an important complement to previous work.

At the implementational level, two neuromodulatory systems may drive these decision dynamics. Previous computational work has proposed that competition between and within corticobasal ganglia-thalamic (CBGT) populations representing specific actions should encode decision uncertainty with explicit links to changes in underlying decision parameters[46,44,43,131]. In addition, the locus coereleus-norepinephrine (LC-NE) system modulates exploration states under uncertainty[10], with changes in firing rates indexing the explore-exploit decision state[79].

Adaptive decision-making is also driven by latent learning. Learning the latent, higher order structure of the environment speeds learning in contexts that share structural similarities, but distinct superficial features[29,83]. Explore-exploit dynamics are linked with this kind of learning under experimental and naturalistic environmental conditions[137,138,62,63], with latent learning of environmental structure shaping future exploration[136]. Understanding how the underlying decision policy evolves while learning higher order structure paints a richer and

more naturalistic picture of the mechanisms driving flexible adaptation.

To shed light on the mechanisms that may be driving dynamic decision policy, this dissertation develops a set of experiments to test how decision policy adapts at two levels of abstraction. As a first step, I examine how decision policies shift choices in response to an environmental change from an algorithmic perspective (Chapter 2). Then I test theoretically motivated predictions about how corticostriatal and locus coeruleus-norepinephrine dynamics may mediate this process (Chapters 2 and 3). Finally, to explore how decision policy reconfiguration may facilitate higher-level inference, I test how these adaptive exploration-exploitation dynamics contribute to learning serially dependent, higher-order structure (Chapter 4).

Contents

Listing of figures

To my mother – for imparting your enduring strength, resilience, determination, analytical streak, and sense of justice.

Acknowledgments

First of all, thanks to my advisor, Tim Verstynen. You taught me to see the world as continuous, with all of the scientific and philosophical implications this brings. You believed in me, *especially* when I didn't. Your patience is superhuman. There isn't enough space here to fully describe the impact of your mentorship and your kinship.

This dissertation would have been impossible without Jeanean Naqvi, Patience Stevens, and Phoebe Dinh. I am eternally grateful for their emotional support, camaraderie, and solidarity. Jeanean – you make me feel loved, accepted, and understood without saying a word and you know how to make me laugh in the most stressful of situations. Patience – your fresh perspective and your balanced view of the gray in any situation renews my sense of wonder. Phoebe – the way you see the layers within any given idea inspires me and I consult my mental simulation of you when I encounter new ones.

To Reggie Raye, for both seeing in me more than I ever had and perpetually challenging my thoughts. For being my interactive theory notebook and a mutual journal of experience. For holding me when I was at my lowest and for reinterpretation. For embracing growth. We are inevitable.

To Chris Wordingham, forever a kindred spirit without limits on his curiosity, his affection, or his loyalty. The meaning of our formative experiences together is ineffable.

To Cassie Eng – you were always there for me when I needed you and you taught me how to be a boss.

To Tim Nolan for helping me parse hard times and for gently nudging me toward social grace. Your capacity to see multiple perspectives helped me navigate my decision trees in the best possible way.

Chapter 1

Introduction

Consider a novice archer attempting to hit a target. Learning to aim for the center of the target and shoot is easy enough for her to learn under stable conditions. But when the wind shifts the trajectory of her arrow, she may not know which set of actions to take to hit the target. With practice she learns to adapt by aiming her arrow opposite to the wind with sufficient magnitude. However, if she wants to adapt under variable conditions, she needs to learn how to adapt her responses without specific experience. Importantly, even if the optimal strategy were given in each scenario, the same intended action may not yield the same reward. This inherent variability of experience comes not only from the dynamics of the environment we find ourselves in, but also from our own behavior[170] and from internal algorithms, both cognitive[52] and implementational[143,152]. Dynamic adaptation would seem to be a difficult problem, then, because we can't possibly have experience with all possible states of the environment (and ourselves).

However, injecting noise into a system can be advantageous. For example, sensory noise in both sensory signals and sensory receptors limits the amount of information available to the central nervous system, acting as a useful constraint on neural computation[49]. Further, behavioral noise can ensure that agents aren't trapped in local minima, preventing higher order learning[169]. More abstractly, signal processing in general can benefit from noise. For example, the concept of stochastic resonance, originally from statistical physics, describes how adding noise to a periodic signal actually *enhances* information transfer for weak signals when the input-output system is nonlinear[58], and this idea has been successfully applied to neural systems under the guise of "stochastic facilitation"[98].

Returning to our archer, absent an explicitly known way to grapple with the influence of these sources of variance on behavior under changing conditions, another way of navigating this problem is to embrace noise. Thus one way the archer could improve her performance is to adopt a generally exploratory posture, sampling the space of possible strategies and observing outcomes. Much research has shown that generally exploring the environment seems to be a good approach to promote flexible learning[20].

However, a more systematic way to navigate shifting environments is to develop a set of rules for when to explore alternative strategies or exploit what is known. In other words, one needs to establish a *decision policy*. But what form should this decision policy take in order to improve the chances of success under

changing conditions?

Emerging research shows that exploration can be dissociated into two general categories – as a bias to acquire information ('directed exploration') and the pure randomization of choice independent of informational value ('random exploration'). Here, both directed and random exploration improve learning[169].

In directed exploration, exploration is pushed toward more informative options using a deterministic information bonus, increasing the value of more informative options. Within a reinforcement learning framework, this can be written as:

$$Q(a) = r(a) + IB(a) \tag{1.1}$$

where $r(a)$ is the expectation of reward for a given choice a and $IB(a)$ represents the 'information bonus' for that action, resulting in a total value for the decision $Q(a)$. Under this framework, if the learner is in a directed exploration state, the choice made will be the one with the highest total value according to this formulation.

In contrast, random exploration drives exploration using noise. This can be written as:

$$Q(a) = r(a) + \eta(a) \tag{1.2}$$

where $\eta(a)$ is zero-mean random noise. Importantly, the exact nature of random exploration is largely dependent on the nature of the noise distribution.

These two forms of exploration are not mutually exclusive[169], and the balance between directed and random exploration may shift according to task demands. Wilson and colleagues put forth the combined formulation:

$$Q(a) = r(a) + IB(a) + \eta(a) \tag{1.3}$$

Here, directed and random influences on exploration have additive effects on the total value of a given choice, with the learner prompted toward exploration when information bonus reaches a threshold or random decision noise pushes the learner away from exploitation.

One approach is to balance these exploration strategies over time, according to varying task demands. If our archer is just beginning to learn, she may not have a pre-existing bias governing which actions are high in informational value. Here, you might expect that the balance tips toward the randomized side of exploration. After practice, she figures out a general approach to aiming her arrow, and realizes which actions might contain the most informational value, biasing her exploration toward the directed form.

How exactly the balance between directed and random exploration evolves over time, and how exactly decision policy in total (including exploitative dynamics) is an open question. One relevant factor driving these dynamics might be the evaluation of evidence over time.

Learning these higher order relationships requires the evaluation of evidence to update decisions in response to environmental change – that is, over time. And, often, the accuracy of our decisions trades off with their speed.

Sequential sampling models capture this tradeoff, and operate under the assumption that noisy information is sampled over time until a threshold of evidence is reached to make a decision[146]. For sequential sampling models measuring the relative degree of evidence for two choices, a response is initiated when the difference in the evidence accumulated for two options reaches a prespecified criterion. When evidence accumulation is discrete, this is known as a random walk model. When evidence accumulation is continuous, this is known as a diffusion process[55]. Within this class, the drift-diffusion model (DDM)[126] assumes that information is continuously integrated over tim (Fig. 1.1). The DDM is one of the most popular accumulation-to-bound models, and it has enjoyed a fruitful history of revealing the processes underlying perceptual and cognitive decisions[128]). The parameters of this model have distinct influences on evidence accumulation, with the drift rate (v) representing the rate of evidence accumulation, the boundary height (a) as the amount of information required to cross the decision threshold, nondecision time (t) as motor-induced delays in the onset of the accumulation process, and starting bias (z) as a bias to begin accumulating evidence for one choice over another.

The parameters of this model exhibit sensitivity to feedback and choice his-

tory[119,127,45,46,99,154]. Specifically, fluctuations in the rate of evidence accumulation (drift-rate) track the relative value of an action[46,100,16,130,26] and fluctuations in the amount of evidence needed to gate a decision (boundary height) track an internal estimate of environmental change[112,172,111,19,26], suggesting that this model is responsive to changes in value estimation. Moreover, new evidence has shown that changes in the drift rate and the boundary height promote adaptive variability when a decision needs to be updated to compensate for a shift in action-outcomes[26], suggesting that such accumulation models are capable of acting as a useful approximation of the processes driving reward-based exploration. Finally, the DDM has the capacity to shape the variability of decisions on a trial-wise basis, allowing us to see both how shifts in parameters change decision policy and sculpt choices and reaction times.

Next, I review two plausible neuromodulatory systems that may modulate decision policy dynamics.

1.3 Implementational mechanisms

1.3.1 The corticobasal ganglia-thalamic (CBGT) circuit.

The corticobasal ganglia-thalamic (CBGT) circuits (Fig. 1.2) are critical for action selection[88]. The canonical model of the CBGT circuit has three major pathways, each serving to suppress (the indirect pathway), promote (the direct pathway), or reactively brake action outputs (the hyperdirect pathway). This canonical model is organized according to multiple action channels[101,23], or sets of pathways associated with a given action (e.g. Left or Right button press),

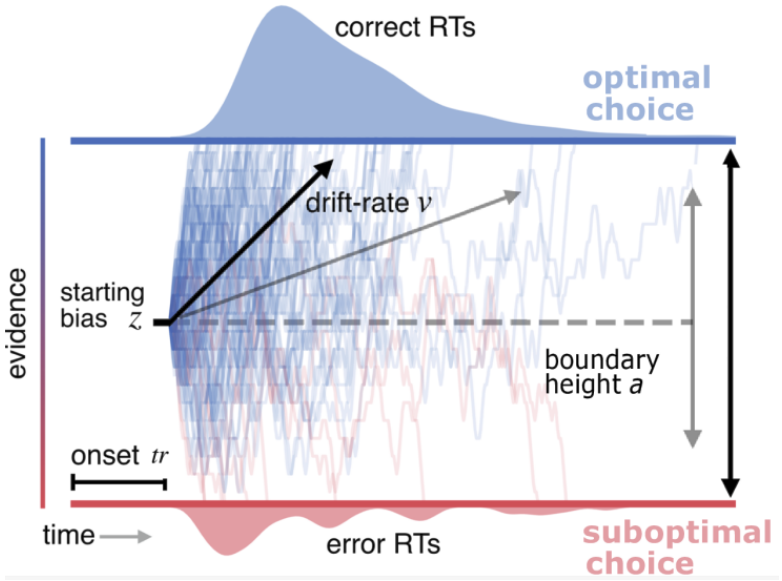


Figure 1.1: The drift diffusion model. The rate of evidence accumulation is the drift rate (v), the amount of information needed to make a decision is the boundary height (a), the starting bias for the decision process is (z), and the non-decision time related to motoric processes is the onset time (tr).

and each action channel contains a direct and an indirect pathway. The division of the CBGT into action channels is supported by empirical findings showing that the execution of specific actions correlates with co-activation of spatially clustered populations of direct and indirect medium spiny neurons (MSNs)[86]. Activation of the direct pathway suppresses the globus pallidus internal segment (GPi), relieving the thalamus from tonic inhibition and allowing it to facilitate action execution by activating primary motor cortex. On the other hand, activation of the indirect pathway activates the globus pallidus external segment (GPe) and subthalamic nucleus (STN) to promote GPi/SNr output, suppressing action.

In addition, the direct and indirect pathways are modulated by dopaminer-

gic feedback from the substantia nigra pars compacta (SNc) during value-based decision-making[75]. Due to the opposing effects of dopamine (DA) on the direct and indirect pathways[139,36], these dopaminergic feedback signals ultimately reinforce rewarded actions and suppress punished actions, suggesting these dynamics should shape value-based decision-making[70,158].

Further, both theoretical[23,25,127,44,46] and experimental[174] evidence suggests that the CBGT circuits are critical to the evidence accumulation process. The topological encoding of actions in the striatum[110,91,1] and the convergence of projections to the GPi/SNr[85,1] suggests that the direct and indirect pathways may compete for control over the output of the basal ganglia, encoding the "evidence" for a given decision as the relative activation of the direct and indirect pathways within the corresponding action channel[16,43]. Critically, this competition between the pathways has been theoretically linked to the rate of information accumulation during decision making[157].

1.3.2 The locus coereleus norepinephrine (LC-NE) system

The locus coeruleus (LC) is a small nucleus in the brainstem, and the main source of the neuromodulator norepinephrine (NE). The LC receives input from a diffuse set of brain regions, including the forebrain, cerebellum, and the brainstem[10]. The LC-NE system has two distinct modes[9] that map onto distinct decision states[10]. In the phasic mode, a burst of LC activity results in a global, temporally precise release of NE. This increases the gain on cortical processing and encourages exploitation. In the tonic mode, NE is released without the tem-

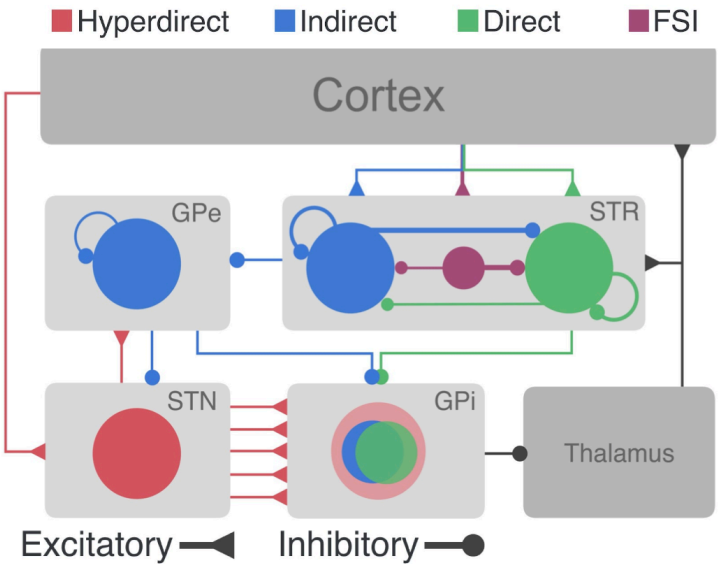


Figure 1.2: The cortico-basal ganglia-thalamic network. The hyperdirect pathway is marked in red, the indirect pathway is marked in blue, and the direct pathway is marked in green. FSI, fast spiking interneurons; GPe, globus pallidus external segment; GPi, globus pallidus internal segment; SNc, substantia nigra pars compacta; STN, subthalamic nucleus; STR, striatum. Diagram adapted[44]

poral precision of the phasic mode, increasing baseline NE[9]. This encourages disengagement from the current task and facilitates exploration. The dynamic fluctuation of these two modes is thought to optimize the trade-off between the exploitation of stable sources of reward and the exploration of potentially better options[10]. Similar to the classic Yerkes-Dodson curve relating arousal to performance[175], performance is optimal when tonic LC activity is moderate and phasic LC activity increases following a goal-related stimulus[11]. Thus the LC-NE system, which can be indirectly measured by fluctuations in pupil diameter[10,79], may be a central mechanism for modulating selection policies.

1.4 Cognitive maps and latent learning

Decision policy is also linked with latent learning under experimental and naturalistic environmental conditions[136,137,138,62,63]. When Tolman introduced the concept of cognitive maps[151], he introduced the concept of latent learning, particularly in terms of latent relational representations that go beyond stimulus-response association. The bulk of research citing Tolman's effects and his theoretical interpretation have been justifiably used to support the idea of spatial maps[115] and their corresponding neural representations of space[47]. Recent findings suggesting that the neural encoding of spatial maps also represent nonspatial features[37,8] have prompted the re-examination of Tolman's cognitive maps to study relational structure at the level of knowledge organization for nonspatial inference in both humans and machines[164,18], with empirical support for the idea that the reorganization of knowledge in terms of cognitive maps aids generalization to shared knowledge structures[96,92,82].

In the reinforcement learning context, the concept of cognitive maps encoding relational structure of the environment has recently re-emerged as an updated version[105] of the successor representation[38]. The successor representation (SR) is a reinforcement learning algorithm that builds a predictive map of the environment to summarize the relationship between states separated by multiple state transitions. To accomplish this long-range prediction of state, the SR occupies an intermediate position on the model-based to model-free continuum of reinforcement learning, balancing the tradeoff between biased and flexible

decision-making[60]:

Standard model-free reinforcement learning updated using temporal difference learning:

$$V(S) = V(S) + \alpha(R_{observed} + \gamma V(s_{new} - V(S)) \tag{1.4}$$

Unlike standard temporal difference learning algorithms that operate over prediction errors in value (Eqn. 1.4), the SR can be learned via a form of temporal difference learning using the difference between observed and predicted state occupancy as the error signal[102] (Eqn. 1.5):

$$M(S) = M(S) + \alpha(onehot(s_{new}) + \gamma M(s_{new} - M(S)) \tag{1.5}$$

Here M represents the successor representation matrix. The $onehot(s_{new})$ keeps track of state visitation. When the agent visits a new state, one visit is added to the count of visits to that state in the row corresponding to that state in M. The sucessor prediction error is the difference between the expected successor of state s from predictive horizon discounted sucessors of the new state. A learning rate α applies to the prediction error.

Offline replay, a memory process in which the hippocampal network internally generates patterns of activation representing compressed versions of prior experience[144], has been suggested to combine current experience with previous memories[120] to guide future behavior[104,103]. Offline replay is not solely a repetition of the past, but a dynamic process sensitive to goal-specification[121] that reverses in response to prediction error[7]. Specifically, human and animal stud-

ies have shown a role for offline replay in inferring latent environmental structure[104,173]. Combining SR with a family of reinforcement learning algorithms called Dyna[147] (SR-Dyna) shows promise as a computational framework for learning relational structure[132]. Here, the predictive map learned by the SR is learned online and state transitions are replayed offline. Mounting evidence supports the plausibility of SR-Dyna in both humans and rats as a computational basis for reinforcement learning[41,105].

Finally, the successor representation can be decomposed into successor features, which abstract successor representations from their context to define primitive components of state representation[102]. This decomposition allows the agent to generalize to tasks that require similar component features[95,102]. As mentioned in the previous section, the complexity of the hierarchical reinforcement learning problem can be drastically reduced by defining subgoals, or "options"[145]. Similarly, decomposing successor features compactly represents abstracted subroutines to reduce the complexity of the problem space while maintaining a state-based representation, and, due to the recombinant nature of these features, this also increases the span of tasks to which the successor representation can generalize because multiple task solutions can be represented as the linear combination of features[95,94].

How exactly humans flexibly balance the adaptive value of noise (exploration) with the value of acting on what they know (exploitation) to dynamically adapt to changing conditions is under-explored. In this dissertation, I first examine how decision policies shift choices in response to an environmental change (Chap-

ter 2). A dynamic variant of the two-armed bandit task[39] designed to impose periodic changes in the most rewarding choice will be used to conduct an initial, exhaustive exploration of how decision policy evolves in response to a change in action-outcome contingencies.

Then I conduct two experiments to test the role of two plausible neuromodulatory systems in these decision policy parameter shifts (Chapters 2 and 3). In Chapter 2, I first consider the locus coereleus-norepinephrine (LC-NE) system, which is known to modulate exploration states under uncertainty. Pupil diameter shows a tight correspondence with LC neuron firing rate[10], with changes in pupillary signal indexing the explore-exploit decision state[79]. Because of this link between LC-NE and the regulation of behavioral variability in response to uncertainty, LC-NE system responses, as recorded by pupil diameter, should associate with the trajectory through decision policy space following a change in action-outcome contingencies. Specifically, if the LC-NE system were sensitive to change points, then I should observe phasic activity following a change in action-outcome contingencies.

In Chapter 3, I test how corticostriatal dynamics might modulate decision policy. Corticostriatal activation is robustly linked to action selection, with direct and indirect pathway activation facilitating and suppressing action selection, respectively[88]. Further, previous work using a biologically realistic spiking CBGT network has shown links between changes in the decision parameters used to define the decision policy I observed in Chapter 2, and CBGT dynamics during learning. In these studies, the rate of evidence accumulation relies

on differences in the simulated ratio of direct pathway spiny projection neurons (dSPNs) to indirect pathway spiny projection neurons (iSPNs) in opposing action channels (left or right selection), while the amount of information needed to make a decision relies on overall iSPN activation across action channels[46,131]. Given the limited resolution of fMRI, I hypothesize this link will largely manifest as a correlation between neural activation associated with the competition between actions and drift rate. Specifically, our quantification of decision uncertainty derived from our decoded neural representation of action selection competition should negatively correlate with the drift rate following a shift in reward contingencies. When drift rate drops after the detection of a change, decision uncertainty should peak. As drift rate recovers to baseline levels, decision uncertainty should be minimized. The nuclei of the CBGT circuit should be essential to this relationship, with a "lesioned" analysis (excluding CBGT activity) failing to show this relationship to the same degree.

Finally, Chapter 4 explores how decision policy adapts in response to latent environmental structure abstracted from any single learned stimulus-response mapping. I test this using a foraging experiment that manipulates the optimally rewarding path in a grid-based world. After learning the initial path, the optimal path is rotated such that the learner needs to exploit second-order knowledge regarding the shape of the path to earn reward. The rotation of the path will test the degree to which participants learn the higher order structure of associations between stimuli. Speeded learning of the rotated path relative to the initially learned path would be evidence of learning the initial higher order

structure. In addition, I predict that behavioral variability (i.e., RT variance and choice variability) should expand after the rotation and contract as participants relearn the adjusted "grid". I expect that participants will follow a decision policy profile similar to those predicted in Chapters 2 and 3, initially starting in a slow exploratory state, with decreased rate of evidence accumulation and, possibly, a transient increase in boundary height. This should be followed by a gradual recovery of drift rate to baseline as the path is found.

Chapter 2

Decision policy dynamically reconfigures under uncertainty

The following text was adapted from Bond, Dunovan, Porter, Rubin, and Verstynen 2021.

2.1 Introduction

"Should I stay or should I go?" refers not only to an iconic 1980s punk anthem but also the fundamental dilemma all animals face in uncertain or unstable environments. Should someone buy coffee from the cafe that serves their favorite roast or try the new cafe that opened down the street? If their favorite drink is bitter one day, is that a sign to switch to a new blend or is one subpar experience inadequate to prompt a switch? Ultimately, these decisions converge to a single predicament: whether we choose an action that we believe is likely to produce desirable results (i.e., exploit) or risk choosing another action that is less certain, on the chance that it will produce a more positive outcome (i.e., explore)[116]. Ultimately, this is the problem of knowing when to change your mind.

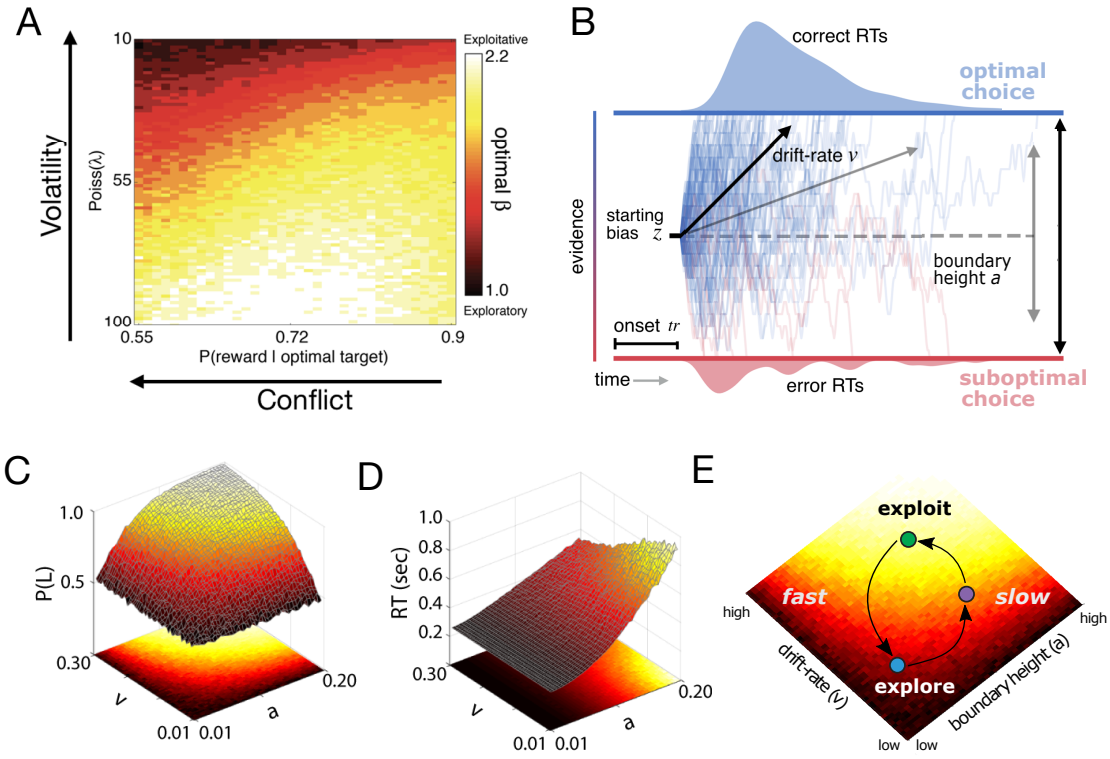


Figure 2.1: Dynamic decision policy reconfiguration. A) The degree of conflict and volatility shifts the optimal balance between exploration and exploitation. B) The drift diffusion model. C) Accuracy (probability that left choice selected is selected; P(L)) as a function of coordinated changes in the rate of evidence accumulation (v) and the amount of information needed to make a decision, or the boundary height (a). D) Reaction time as a function of changes in the rate of evidence accumulation and the boundary height. E) Decision policy reconfiguration.

The shift of a decision policy from exploratory to exploitative states is driven by environmental context. To illustrate this, Figure 2.1A shows what happens when a simple reinforcement learning (RL) agent tries to maximize reward in a dynamic variant of the two-armed bandit task ([148]; see Methods). Here, the relative difference in reward probability for the two actions (conflict) and the frequency of a change in the optimal action (volatility) were independently ma-

nipulated. For each level of conflict and volatility, a set of tabular Q-learning[148] agents played the task with learning rate held constant while the degree of randomness of the selection policy (β in a Softmax function) varied. The agent that returned the most rewards was identified as the agent with the best exploration-exploitation balance. Increasing either form of uncertainty led to selecting agents with more random or exploratory selection policies (Fig. 2.1A). As the value of the optimal choice decreases relative to the value of a suboptimal choice (conflict increases), the learner exploits what she already knows. Action values grow unstable (volatility increases) when the clarity of the optimal choice is constant (constant conflict), and the learner is biased toward exploration[21]. As these two forms of uncertainty change together, the gradient of action selection strategy also changes.

Knowing *how* decision policies shift in the face of dynamic environments requires looking at the algorithmic properties of the policy itself. One popular set of algorithms for describing the dynamics of decision making are accumulation-to-bound processes like the drift-diffusion model (DDM[126]). The normative form of the DDM proposes that a decision between two choices is described by a noisy accumulation process that drifts towards one of two decision boundaries at a specific rate (Fig. 2.1B). Two parameters of this model are critical in determining the degree of randomness of a selection policy: the rate of evidence accumulation (drift rate; v) and the amount of information required to make a decision (boundary height; a). For example, decreasing the drift rate and increasing the boundary height leads to more random decisions (Fig. 2.1C), with

the speed of these decisions depending on the ratio of the two parameters (Fig. 2.1D). Thus exploratory policies can result in either fast or slow decisions, depending on the relative configuration of drift rate and boundary height.

Are the parameters that govern accumulation of evidence for decision making modifiable? Previous modeling work has shown that the parameters of a DDM process can be modulated by feedback signals and choice history[119,127,45,46,99,154] with different mechanisms for adapting the drift rate and the boundary height. In value-based decision-making tasks where the statistics of sensory signals are equivalent for all actions, drift rate fluctuations appear to track the relative value of an action or the value difference between actions[46,100,16,130]. In contrast to value estimation, selection errors in this context have been linked to changes in the boundary height[54,53,25,72,71,46,45] and internal estimates of environmental change[112,172,111,19].

One plausible neural mechanism for this migration along the surface of selection policies is the locus coereleus norepinephrine (LC-NE) system, which has been linked to adaptive behavioral variability in response to uncertainty[153,40,28]. The LC-NE system has two distinct modes[9] that map onto distinct decision states[10]. In the phasic mode, a burst of LC activity results in a global, temporally precise release of NE. This increases the gain on cortical processing and encourages exploitation. In the tonic mode, NE is released without the temporal precision of the phasic mode, increasing baseline NE[9]. This encourages disengagement from the current task and facilitates exploration. The dynamic fluctuation of these two modes is thought to optimize the trade-off between the

exploitation of stable sources of reward and the exploration of potentially better options[10]. Thus the LC-NE system, which can be indirectly measured by fluctuations in pupil diameter[10,79], may be a central mechanism for modulating selection policies.

We investigated the malleability of decision policies as the environment necessitates a change of mind as to what constitutes the "best" decision. To control environmental uncertainty, we manipulated the volatility of changes in action-outcome contingencies (i.e., which of two targets returns the most rewards), as well as ambiguity in optimal choice (*conflict*), while human participants performed a dynamic variant of the two-armed bandit task[149]. We predicted that, in response to suspected changes in action-outcome contingencies, humans would exhibit a stereotyped adjustment in the drift rate and boundary height that pushes decisions from certain, exploitative states to uncertain, exploratory states and back again (Fig. 2.1E). In addition, using pupillary data, we explored whether the LC-NE system covaries with shifts of the boundary height in response to a change in action outcomes to facilitate exploration, consistent with prior studies[84,108,34].

2.2 Results

Across two experiments we used a dynamic two-armed bandit task with equivalent sensory reliability across arms to independently manipulate the reward conflict and the volatility of action outcomes in order to measure how underlying decision processes respond to changes in action-outcome contingencies (see

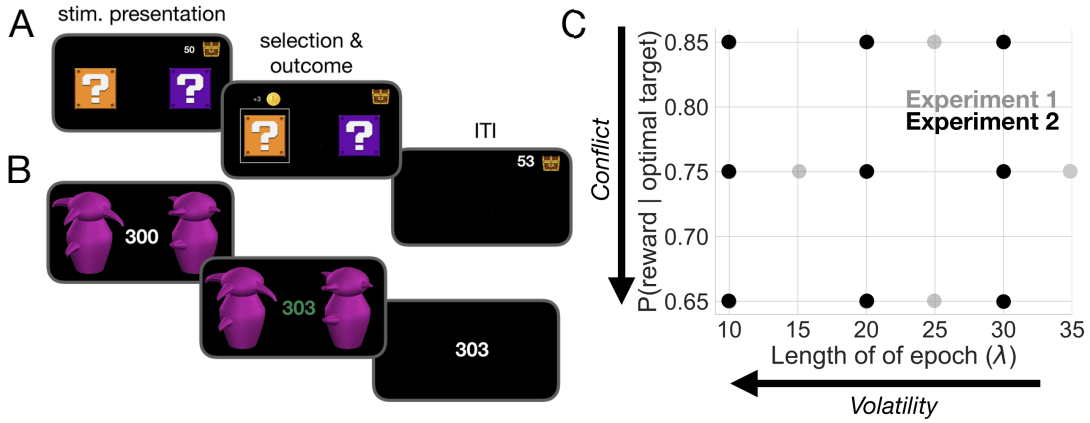


Figure 2.2: Task and uncertainty manipulation. A) In Experiment 1, participants were asked to choose between one of two "mystery boxes". The point value associated with a selection was displayed above the chosen mystery box. The sum of points earned across trials was shown to the left of a treasure box on the upper right portion of the screen. B) In Experiment 2, participants were asked to choose between one of two Greebles (one male, one female). The total number of points earned was displayed at the center of the screen. The stimulus display was rendered isoluminant throughout the task. C) The manipulation of conflict and volatility for Experiments 1 (gray) and 2 (black). Each point represents the combination of degrees of conflict and volatility. Under high conflict, the probability of reward for the optimal and suboptimal target is relatively close. Under high volatility, a switch in the identity of the optimal target selection is relatively frequent.

Stimuli and Procedure). Both of these experiments shared a common feedback structure. Participants were asked to select either the left or right target presented on the screen using the corresponding key on a response box. Rewards were probabilistically determined for each target and, if a reward was delivered, it was sampled from a Gaussian distribution. The optimally rewarding target delivered reward with a predetermined probability ($P(optimal)$) and the suboptimal target gave reward with the inverse probability ($1 - P(optimal)$). After a delay determined by the rate parameter of a Poisson distribution (λ), the reward probabilities for the optimal and suboptimal targets would switch.

In Experiment 1, twenty-four participants completed four sessions (high and low conflict; high and low volatility) each composed of 600 trials. During each session, they were asked to select one of two coin boxes (Exp. 1: Fig. 2.2A). The levels of conflict and volatility for all four conditions in Experiment 1 are shown as gray dots in Fig. 2.2C. Experiment 2 was a replication of Experiment 1 with more extensive within-subject sampling of conflict and volatility, as well as the inclusion of pupilometry as a proxy for measuring LC-NE dynamics. In Experiment 2, participants were asked to choose between one of two Greebles (one male, one female). Each Greeble probabilistically delivered a monetary reward (Exp. 2: Fig. 2.2B). Participants were trained to discriminate between male and female Greebles prior to testing to prevent errors in perceptual discrimination from interfering with selection on the basis of value estimation. Four participants completed nine sessions composed of 400 trials each, generating 3600 trials in total per subject. The levels of conflict and volatility for all

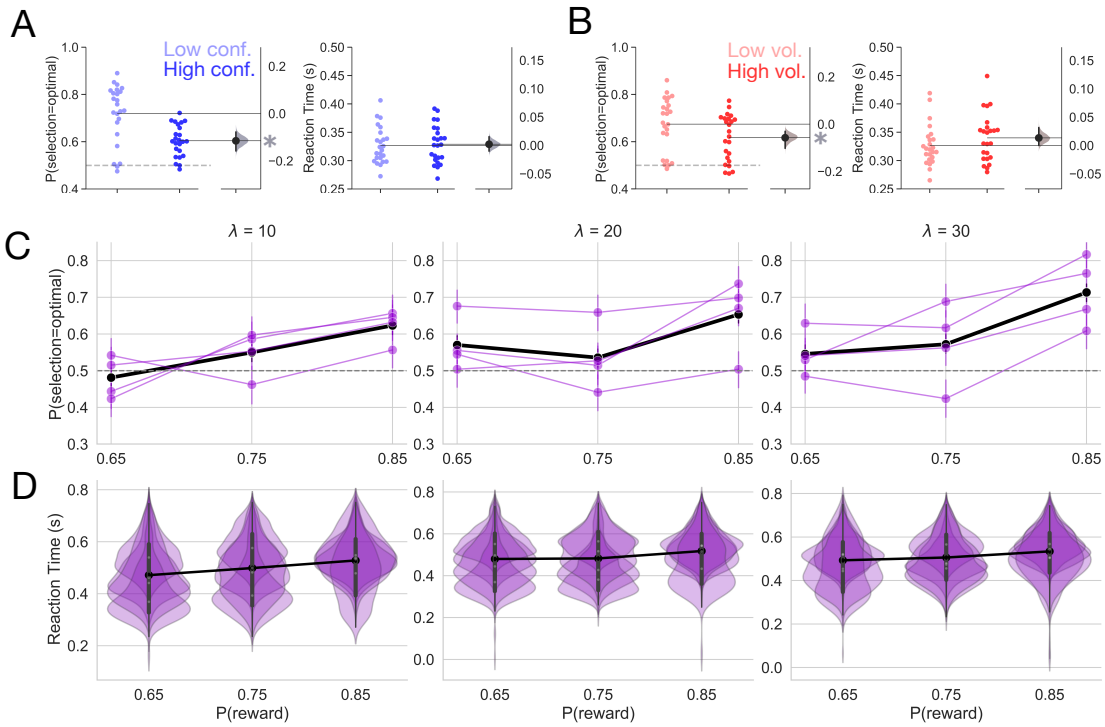


Figure 2.3: Behavior. A) Mean accuracy and reaction time for the manipulation of conflict in Experiment 1. B) Mean accuracy and reaction time for the manipulation of volatility in Experiment 1. Each point represents the average for a single subject. The distribution to the right represents the bootstrapped uncertainty in the mean difference between conditions (high conflict or high volatility subtracted from low conflict or low volatility). Distributions with 95% CIs that do not encompass 0 are marked with an asterisk. C) Mean accuracy for Experiment 2. Each purple line represents a subject. The black line represents the mean accuracy calculated across subjects. D) Reaction time distributions for each subject for Experiment 2. The black line represents the mean reaction time calculated over subjects. Error bars indicate a bootstrapped 95% confidence interval. For panels C and D, λ values shown above each plot specify the average period of optimal choice stability and the probability of reward shown on the x-axis specifies the degree of conflict. Means are calculated over all trials.

nine conditions in Experiment 2 are shown as black dots in Fig. 2.2C. Importantly, Experiment 2 manipulated the same forms of uncertainty as Experiment 1, but had different perceptual features and more expansively sampled the space of conflict and volatility. Given the similarity in design, the behavioral results for both of these experiments are presented together below.

2.2.1 The influence of ambiguity and instability on speed and accuracy

We first looked at overall speed and accuracy effects in both Experiments 1 and 2. In Experiment 1, accuracy (i.e., optimal choice selection) suffered as the optimal choice grew more ambiguous, with accuracy in the low conflict condition being 1.2 times higher than what is observed in the high conflict condition (Fig. 2.3A; $\beta = 1.213$, 95% CI: 1.192, 1.235, z=21.36, p<2e-16). In contrast, increasing conflict had no observable impact on overall reaction times (Fig. 2.3A; $\beta = -6.902e^{-}5$, 95% CI: -0.002, 0.002, t =-0.06, p=0.951). As expected, participants also became less accurate as the instability of action outcomes (i.e. volatility) grew (Fig. 2.3B; $\beta = 0.092$, 95% CI: 0.077, 0.111, z=10.36, p<2e-16). Under volatile conditions, participants also took slightly longer to make a decision ($\beta = -0.012$, 95% CI: -0.015,-.010, t=-10.80, p<2e-16); however, while this effect on reaction times was statistically reliable, the impact of volatility on reaction times was weak (increasing volatility increased reaction time by $\sim$ 13 ms on average; Fig. 2.3B).

Experiment 2 served as a high powered test of whether the effects we observed in Experiment 1 were replicable at the within-subject level. Because Experiment 2 independently manipulated conflict and volatility, we were able to test whether conflict and volatility interacted to affect behavior. We found similar effects of conflict and volatility on accuracy as we observed in Experiment 1 (Fig. 2.3C). Accuracy increased as conflict decreased (i.e. as the probability of reward increased; β=0.223, 95% CI=0.189,0.256, z=12.757, p<2e-

16). As the environment grew less volatile, accuracy increased (β=0.101, 95% CI=0.066,0.14, z=5.828, p=5.6e-09). We did not observe an interaction of conflict and volatility on accuracy (β=0.024, 95% CI=-0.013, 0.058, z=1.364, p=0.173).

However, we did find that conflict and volatility interacted to affect reaction time (RT; β=-0.002, 95% CI=-0.004, -0.001, t=-3.084, p=0.002), with a linear increase in reaction time as the environment grew less volatile and conflict was highest (when $p(r) = 0.65$ $\bar{RT} = 0.472, 0.480, 0.493$ as a function of λ; see Fig. 2.3D for RT distributions). When conflict was moderate ($p(r) = 0.75$) or low ($p(r) = 0.85$), volatility had a nonlinear effect on RTs. Here, reaction times decreased when volatility was moderate ($\bar{RT} = 0.483$ when λ=20 and $p(r) = 0.75$; $\bar{RT} = 0.518$ when λ=20 and $p(r) = 0.85$). Reaction times increased to approximately the same extent within moderate or low conflict conditions when volatility was high ($\bar{RT} = 0.499$ when λ=10 and $p(r) = 75$; $\bar{RT} = 0.528$ when λ=10 and $p(r) = 0.85$) and when volatility was low ($\bar{RT} = 0.506$ when $\lambda = 30$ and $p(r) = 0.75$; $\bar{RT} = 0.534$ when $\lambda = 30$ and $p(r) = 0.85$), with an increase in baseline reaction times when conflict was low relative to moderate ($\bar{RT} = 0.527$ when $p(r) = 0.85$; $\bar{RT} = 0.496$ when $p(r) = 0.75$; see Supp. Fig. **??** for interaction visualization).

At the gross level, over all trials within an experimental condition, increasing the ambiguity of the optimal choice (conflict) and increasing the instability of action outcomes (volatility) decreases the probability of selecting the optimal choice. Reaction time effects were inconsistent, with a negligible effect of volatility in Experiment 1. Experiment 2 revealed that volatility and conflict inter-

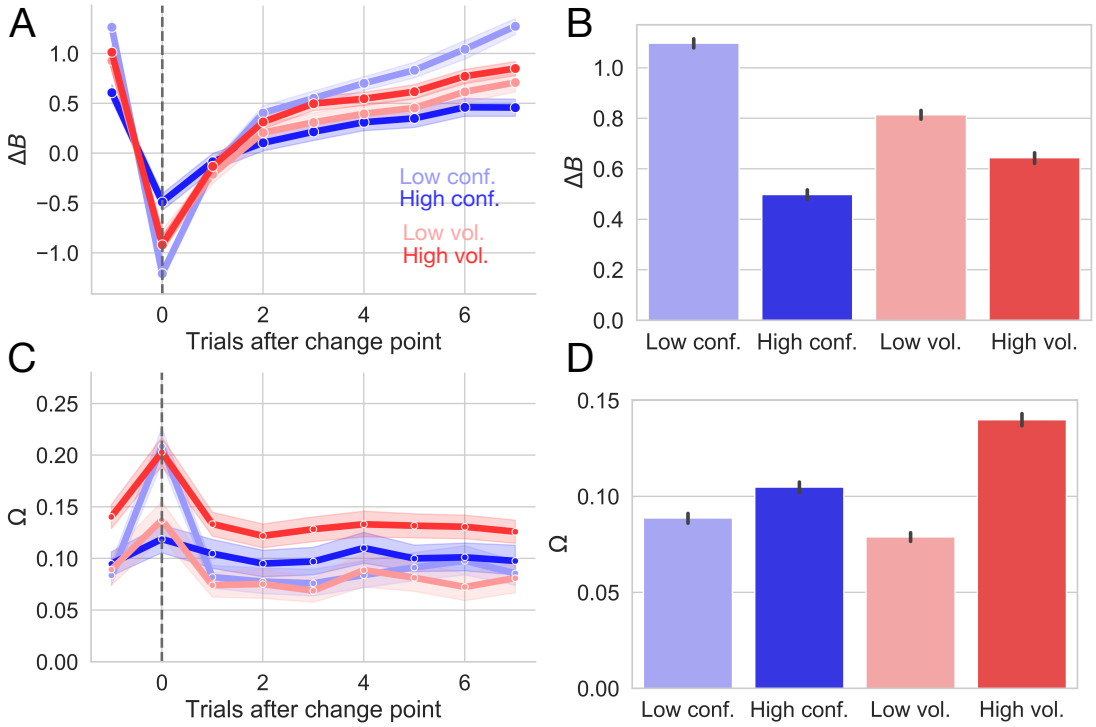


Figure 2.4: Changes in ideal observer estimates as a function of condition for Experiment 1. A) Changes in the belief in the value of the optimal target (ΔB) as a function of conflict and volatility over time. B) Belief in the value of the optimal choice by condition and averaged over all trials. C) Changes in change point probability (Ω) as a function of conflict and volatility over time. D) Change point probability by condition and averaged over all trials. Error bars represent 95% CIs.

act to influence reaction times in complex ways. However, because trials where action-outcome contingencies change are so infrequent, even under high volatility conditions, these overall effects on speed and accuracy may be masking more subtle behavioral dynamics in response to feedback changes. We adopt a more focal, model-based analysis in the next section to clarify these peri-change point dynamics.

2.2.2 Tracking estimates of action value and environmental volatility

We calculated trial-by-trial estimates of two ideal observer parameters of environmental states (see Cognitive model for calculation details;[112,155]). Belief in the value difference (ΔB) reflects the difference between the learned values of the optimal and suboptimal targets. For ease of interpretation, we refer to the converse of belief as doubt, such that when belief decreases doubt increases. ΔB thus reflects a local estimate of uncertainty regarding the choices themselves. To capture the estimated probability of fundamental shifts in action values, we calculated how often the same action gave a different reward (change point probability; Ω). Here Ω reflects a global estimate of uncertainty in the environment, specifically the uncertainty in response contingencies. We used the data from Experiment 1 to assess how well these learning estimates captured our imposed manipulations, and observed similar results in Experiment 2 (Supp. Fig. **??**).

In Experiment 1 we observed a sharp decrease in ΔB after a switch in action outcomes and a gradual return to asymptotic values (Fig. 2.4A) with a decreased difference in reward probability resulting in increased doubt (2.4B; $\beta = 0.216$, 95% CI:0.206, 0.224, t=46.24, p<2e-16).

As expected, less volatile conditions allowed the learner to more fully update her belief in the value of the optimal choice over all trials ($\beta = 0.058$, 95% CI:0.050, 0.068, t=12.32, p<2e-16), though to a smaller degree than low conflict conditions allowed (see Fig. 2.4B). Increasing volatility resulted in a sharp increase in the estimate of Ω at the onset of a change point with a quick return

to a baseline estimate of change (Fig. 2.4C). Notably, this estimate of Ω was more sensitive to change points when conditions were relatively volatile, with a more pronounced peak in response to a change under high volatility conditions than under low volatility conditions (Fig. 2.4C). Correspondingly, over all trials, Ω was higher under more volatile conditions (Fig. 2.4D, $\beta = -0.022$, 95% CI:-0.023, -0.020, t=-30.74, p<2e-16) indicating sensitivity to the increased frequency of action outcome switches in the reward schedule.

When the identity of the optimal choice was clear (i.e. when conflict was low), the estimate of Ω was more sensitive to the presence of a true change point than when the optimal choice was ambiguous (i.e. when conflict was high) (Fig. 2.4C,D). This observation is consistent with the idea that increasing the difficulty of value estimation and, thereby, the assignment of value to a given choice also impairs change point sensitivity. Interestingly, increasing conflict nevertheless resulted in a net increase in Ω calculated over all trials (Fig. 2.4D; β=-0.006, 95% CI:-0.007, -0.004, t=-8.64, p<2e-16), likely because higher conflict conditions increased the baseline estimate of change instead of enhancing sensitivity to true change points (see change point response and relative baseline values for the high conflict condition in Fig. 2.4C). Here, the system conservatively over-estimates the volatility of action outcomes, assuming a slightly greater frequency of changes in the probability of reward for the optimal choice than we imposed (actual proportion of change points for high conflict condition: 0.041 ± 0.004; estimated Ω: 0.105 ± 0.014).

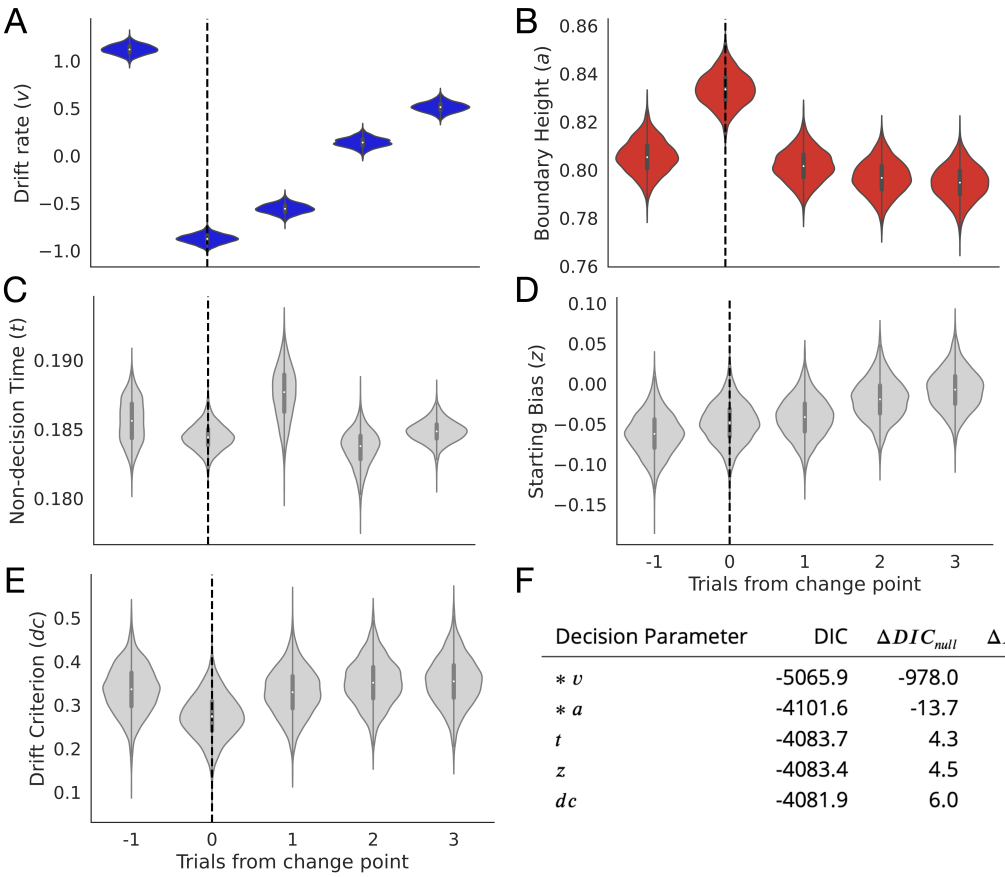
Reassuringly, net change point probability was much greater when change

points were more frequent (see increased Ω estimates for high volatility conditions over high conflict conditions in Figure 2.4D). These results suggest that our formulation of these ideal observer estimates adequately captures our manipulation of volatility and conflict at a continuous level.

Thus, these ideal observer parameters show a reliable response to a change in action-outcome contingencies. The difference in value belief decreases, or doubt increases, when a change point occurs and slowly recovers over the course of six to eight trials as participants learn new action-outcome contingencies. The initial drop in belief difference is deeper and the recovery time after a change point is slower in conditions with greater overall uncertainty (i.e. under high conflict and high volatility). In contrast, internal estimates that a change has occurred briefly spike at a change point, indicating that participants can reliably detect that something has changed, and quickly settle after a few trials. Interestingly, net change point probability estimates are higher in the conditions with higher uncertainty (high conflict, high volatility), likely reflecting increased vigilance for changes in those conditions. In the next section we explore how the underlying parameters of the decision process itself respond to local changes in action-outcome contingencies.

2.2.3 Different forms of uncertainty impact distinct decision processes

Our next goal was to test which decision parameters were sensitive to a change point. To this end, we estimated the change point evoked response of the bound-



F

Decision Parameter	DIC	ΔDIC_{null}	ΔDIC_{best}
$* v$	-5065.9	-978.0	0.0
$* a$	-4101.6	-13.7	964.3
t	-4083.7	4.3	982.2
z	-4083.4	4.5	982.4
dc	-4081.9	6.0	984.0

Figure 2.5: Change point sensitivity of underlying decision processes. Posterior distributions for each decision parameter are shown for the trial prior to a change point to three trials after the change point. A) The drift rate. B) The boundary height. C) Non-decision (onset) time. D) Starting bias. E) Drift criterion. F) Degree of fit to observational data as information loss. The models that lost the least information are marked with an asterisk.

ary height a, drift rate v, non-decision time t, starting bias z, and drift criterion dc for each trial surrounding the change point. To detect changes in the change-point-evoked distributions for each decision parameter, we evaluated whether the sequential distributions evoked by each trial were significantly different, beginning with the trial preceding the change point and ending three trials after

the change point. For example, if the 95% CI of the z distribution evoked on the trial prior to the change point overlapped with the 95% CI of the distribution evoked on the change point and so on for all successive trials considered, then we would conclude that z failed to show change point sensitivity (see Hierarchical drift diffusion modeling for details). To select the model that best accounted for the data, we compared the deviance information criterion (DIC) scores[141] for these models. DIC scores provide a measure of model fit adjusted for model complexity and quantify information loss. A lower DIC score indicates a model that loses less information. Here, a difference of ≤ 2 points from the lowest-scoring model cannot rule out the higher scoring model; a difference of 3 to 7 points suggests that the higher scoring model has considerably less support; and a difference of 10 points suggests essentially no support for the higher scoring model[141,30].

Under this analysis, we found that only the boundary height and drift rate showed change point sensitivity as defined above. The drift rate showed a clear, persistent separation between trial-specific distributions, with a rapid decrease at the onset of the change point (t_{-1} 95% CI = 1.021, 1.218; t_0=-0.972, -0.779) and a return to baseline values thereafter (t_1=-0.656, -0.46; t_2=0.039, 0.241; t_3=0.411, 0.616; Fig. 2.5A). The boundary height showed a transient response to the change point, spiking (t_{-1} 95% CI = 0.792, 0.819; t_0=0.820, 0.847) and then dropping to baseline levels (t_1=0.789, 0.815; t_2=0.783, 0.811; t_3=0.780, 0.808; Fig. 2.5B).

The remainder of the decision parameters showed no change point sensitiv-

ity. Non-decision time showed no clear response (t_{-1} 95% CI = 0.183, 0.188; t_0=0.183, 0.186; t_1=0.184, 0.191; t_2=0.181, 0.186; t_3=0.183, 0.186; Fig. 2.5C) along with the starting bias (t_{-1} 95% CI = -0.112, -0.01; t_0=-0.098, 0.002; t_1=-0.090, 0.008; t_2=-0.069, 0.032; t_3=-0.055, 0.045; Fig. 2.5D) and the drift criterion (t_{-1} 95% CI = 0.229, 0.439; t_0=0.175, 0.374; t_1=0.223, 0.435; t_2=0.245, 0.458; t_3=0.244, 0.464; Fig. 2.5E).

Further, models fitting drift rate and boundary height lost the least null-model-adjusted information relative to models of the change-point-evoked response for the other parameters, showing that a change-point-evoked decrease in drift rate and spike in the boundary height best accounted for our observational data in comparison to all alternatives (ΔDIC_{null} for v = -978 and ΔDIC_{null} for a = -13.7; see Figure 2.5F).

Given that only the drift rate and boundary height showed change point sensitivity, we next focused on how those two parameters related to internal estimates of change and conflict in both experiments. Recall that we used the ideal observer parameters ΔB and Ω as proxies for internal estimates of belief in the difference in learned target values and change point probability, respectively. This provided a continuous quantification of our manipulation of conflict and volatility (see Tracking estimates of action value and environmental volatility). Experiment 2 provided an intensively sampled within-subject test of the change-point-evoked mapping between decision processes and these ideal observer estimates.

In order to determine the nature of the mapping between the ideal observer

parameters and the change-point sensitive decision parameters, we estimated single and dual-parameter models mapping ΔB and Ω and the change-point-sensitive decision parameters, drift rate and boundary height, and examined the fit of these models to our data. We found that the model mapping ΔB to drift rate and Ω to boundary height provided the best fit in Experiment 1 ($\Delta DIC_{null} = -2698.0$; left panel of Table 2.1.

To test whether this mapping was preserved in an independent data set, we performed the same model comparison procedure for Experiment 2. Because Experiment 2 followed a replication-based design, we fit a separate model to each subject to assess the replicability of the best fitting model from Experiment 1. While we found support for the model mapping ΔB to drift rate and Ω to boundary height, we also found that the DIC scores for the single-parameter model mapping ΔB to v alone fit the data equally well (see right panel of Table 2.1 for summary statistics and Supp. Fig. **??** for subject-wise values). Altogether, this suggests that we have strong evidentiary support for a mapping between value-driven belief and drift rate (Fig. 2.6A, blue). However, the support for a mapping between change point probability and boundary height (Fig. 2.6A, red), while robustly present in Experiment 1, fails to appear when tested in an independent data set.

For a more granular assessment of how drift rate and boundary height respond to a change point, we quantified the change-point-evoked effect of ΔB and Ω on drift rate and boundary height , respectively, for both experiments (see Hierarchical drift diffusion modeling for details). In Experiment 1, we found

that the rate of evidence accumulation, v, increased with the belief in the value of the optimal choice relative to a change point ($\beta_{v\sim\Delta B} = 0.576$, 95% CI: 0.544, 0.609, empirical $p = 0.000$; Fig. 2.6B, left panel). The boundary height *increased* with change point probability ($\beta_{a\sim\Omega} = 0.046$, 95% CI: 0.005, 0.088, empirical $p = 0.001$; Fig. 2.6B, right panel).

Experiment 2 showed similar, but attenuated, results, with drift rate increasing with ΔB (ΔB ($\beta_{v\sim\Delta B} = 0.112$, 95% CI: 0.016, 0.227, empirical $p = 0.004$; Fig. 2.6B, inset panel on left) and an unreliable effect of Ω on boundary height ($\beta_{a\sim\Omega} = -0.036$, 95% CI: -0.155, 0.097, empirical $p = 0.282$; Fig. 2.6B, inset panel on right). Therefore, as the belief in the value of the optimal choice approaches the reward value for the optimal choice, the rate of information accumulation increases. An internal estimate of change point probability weakly increases the amount of information required to make a decision, although this latter effect is less reliable.

Altogether, these results suggest a drift rate mechanism for adaptation to change that may also combine with boundary height dynamics (Fig. 2.6A). However, the strength of the drift rate response weakened and the boundary height response was statistically unreliable in Experiment 2 (Fig. 2.6B, inset panels). When a change point is detected and the threshold for committing to a choice (a) responds, it shows a weak, transient increase. At the same time, the drift rate approaches zero, allowing time for the decision process to diffuse and encouraging a random selection. As the learner accrues information about the new optimal choice, the rate of information accumulation slowly recovers to

Experiment 1

	ΔB	Ω	DIC	$\Delta\text{DIC}_{\text{null}}$	$\Delta\text{DIC}_{\text{best}}$
***I**	v	a	-18643.9	-2698.0	0.0
II	a	v	-16265.6	-319.7	2378.3
III	–	v	-16180.5	-234.7	2463.3
IV	v	–	-18630.8	-2684.9	13.1
V	–	a	-15949.20	-3.4	2694.7
VI	a	–	-16032.8	-87.0	2611.1
VII	–	–	-15945.8	0.0	2698.0

Experiment 2

	ΔB	Ω	$\Delta\text{DIC}_{\text{null}}$	$\Delta\text{DIC}_{\text{best}}$
$^{*\sim}$**I**	v	a	-90.3 $\pm$71.7	1.0 $\pm$0.8
II	a	v	-7.6$\pm$13.1	83.8 $\pm$60.5
III	–	v	-8.5 $\pm$13.1	82.9 $\pm$61.4
$^{*\sim}$**IV**	v	–	-90.8 $\pm$71.0	0.5 $\pm$1.1
V	–	a	0.3 $\pm$2.5	91.6 $\pm$70.6
VI	a	–	0.95 $\pm$1.4	92.3 $\pm$70.9
VII	–	–	0 $\pm$0	91.3 $\pm$71.5

Table 2.1: Model comparison for Experiments 1 and 2. Roman numerals refer to a given model, as defined by the mapping between the ideal observer estimates and decision parameters in the first two columns. The left panel shows the deviance information criterion (DIC) scores for the set of models considered during the model selection procedure for Experiment 1. The right panel shows the DIC scores for the equivalent model selection analysis for Experiment 2, with a model estimated for each of four subjects. Values shown represent the mean and standard deviation computed over subjects. Note that the raw DIC values for each of the subjects in Experiment 2 are included in Supplementary Table **??**). The column labeled DIC gives the raw DIC score, ΔDIC_{null} lists the change in model fit from an intercept-only model (the null-adjusted fit), and ΔDIC_{best} provides the change in null-adjusted model fit from the best-fitting model. The best performing model is denoted by an asterisk, with equivocal best cases marked by a tilde.

asymptotic levels, with the decision process assuming a more directed path toward the choice that has accrued evidence for reward. Together, the changes in these underlying decision processes, largely driven by drift rate dynamics, point to a mechanism for gathering information in a relatively slow, unbiased manner shortly after the learner suspects she should update her valuation. We now explore these dynamics in more detail in the next section.

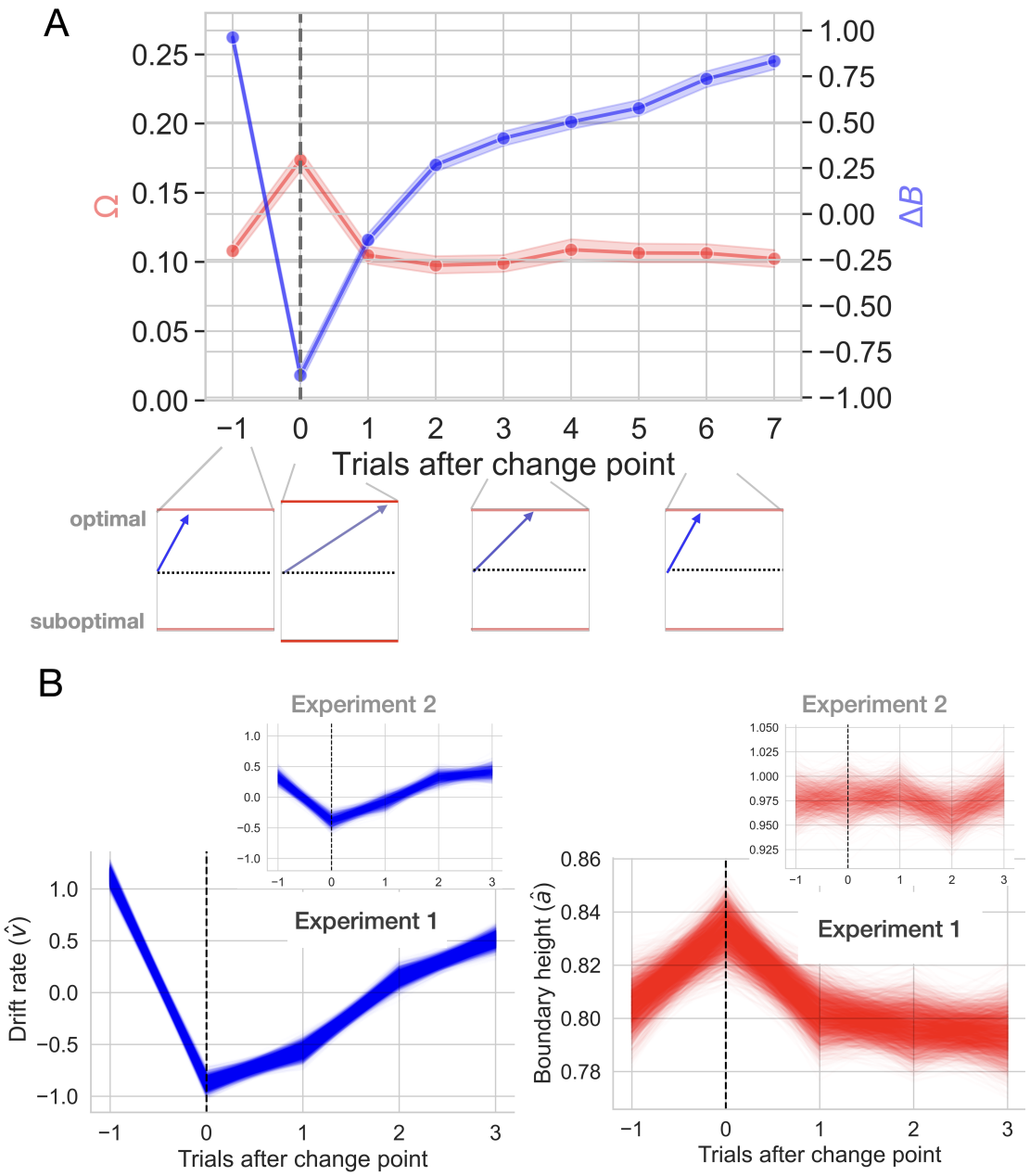


Figure 2.6: Change-point-evoked uncertainty. A) Changes in ideal observer estimates of uncertainty over time and their effect on the boundary height and the drift rate. Directly after a change point, the boundary height *increases* and the drift rate slows. Over time, the boundary height returns to its baseline value and the drift rate increases. B) Fitted estimates of change-point-evoked drift rate and boundary height for both experiments with 95% CIs of the posterior distributions. Inset plots represent data from Experiment 2.

2.2.4 Environmental instability prompts a stereotyped decision trajectory

So far we have established that both the drift rate and the boundary height can be independently manipulated by two different estimates of environmental uncertainty with different temporal dynamics, although this effect reduces to drift rate dynamics in Experiment 2. This suggests that a change in action-outcome contingencies prompts a unique trajectory through the space of possible decision policies (Fig. 2.1E).

To visualize this trajectory, we plot the temporal relationship between drift rate and boundary height beginning with the trial prior to the change point and ending three trials after the change point (Fig. 2.7A). To clearly visualize the distribution of the change-point driven response in the relationship between drift rate and boundary height over time, we also represent the trialwise shift in these two decision variables as vectors. The trial-by-trial estimates of drift rate and boundary height were taken from the best model of the fitted change-point-evoked response and z-scored (see Different forms of uncertainty impact distinct decision processes for model selection). Then the difference between each sequential set of boundary height and drifr rate coordinates, (a, v), was calculated to produce a vector length. The arc tangent between these differenced values was computed to yield an angle in radians between sequential decision vectors, concisely representing the overall decision dynamics (θ, Fig. 2.7B; see Decision vector representation for methodological details).

For Experiment 1, following a shift in response contingencies, the naviga-

tion of this decision surface follows a stereotyped pattern. The boundary height spikes and drift rate decreases rapidly, gradually recovering and stabilizing over time (see the trial prior to the change point in Fig. 2.7A). This decision trajectory is robust in Experiment 1 (Fig. 2.7B, top panel).

Here, we find that the distribution of θ prior to a change point averages to $\sim 300°$, sharply changes in response to the observation of a change point ($\sim 165°$) and steadily returns to values prior to the onset of a change (main panels in Fig. 2.7B). One trial after the change point, drift rate sharply decreases and boundary height spikes, after which boundary height quickly recovers and drift rate steadily progresses toward its baseline value.

However, this trajectory is substantially more variable in Experiment 2, with most of the response restricted to the drift rate dimension and inconsistent trajectories along the boundary height dimension (Fig. 2.7B, lower panel). Here, the distribution of θ prior to a change point averages to $\sim 270°$ and shifts to $\sim 90°$ with the observation of a change. In both experiments, we find that the decision trajectory quickly responds to a shift in action outcomes and also quickly recovers and stabilizes.

Having characterized the change-point-evoked trajectory through the range of decision policies, we next asked whether conditions of increased volatility and increased conflict might modify its path. To this end, we conducted a comparison of a null model with models specifying the change-point evoked response alone and this evoked response as a function of conflict and volatility. To estimate this relationship between drift rate and boundary height, we used

Bayesian circular regression[106]. First, we tested the null hypothesis that the decision dynamics (the relationship between drift rate and boundary height; θ) were solely a function of the intercept, or the average of the decision dynamics θ:

$$\theta = \beta_0$$

We call this the null model.

To test the hypothesis that decision dynamics varied solely as a function of time after a switch in action-outcome contingencies, we estimated the vector of change in (a, v) coordinates (θ) relative to a change point, with the time scale of consideration determined by the results of a stability analysis from Experiment 1 (see Model proposals and evaluation; Supp. Fig. **??**):

$$\theta = \beta_0 + \beta_{\Delta t_{i:3}}$$

We call this the evoked response model.

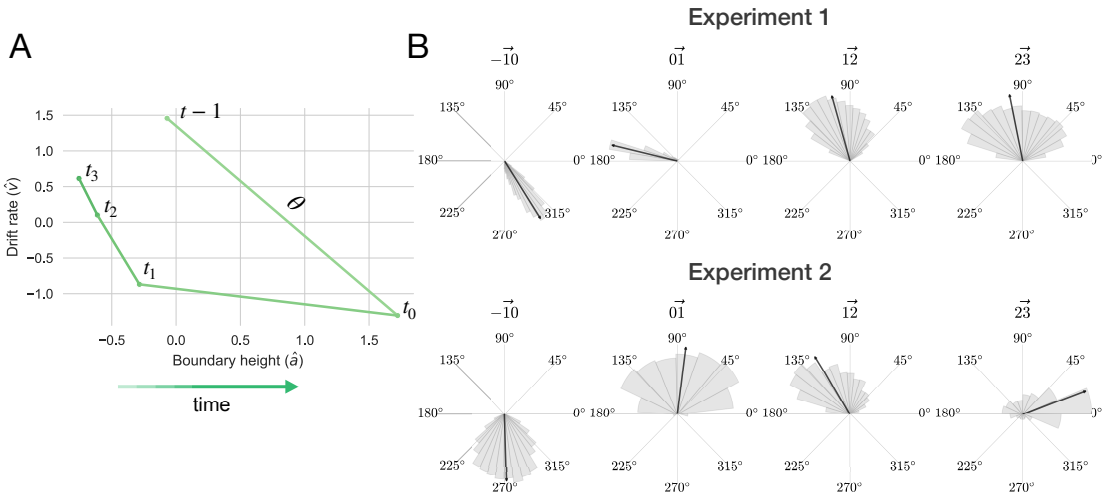


Figure 2.7: The decision surface. A) Representing decision space in vector form. An angle (θ) was calculated between sequential values of (a,v) coordinates, beginning with the trial prior to the change point. This represents subject-averaged data from Experiment 1. Note that these trajectories are z-scored. B) Distributions depicting the angle between drift rate and boundary height for both Experiments 1 and 2. Each subpanel shows the distribution of angles between (a, v) over sequential trials, beginning with the trial prior to the change point. The area of the shaded region is proportional to the density and the arrow represents the circular mean.

Our model comparison logic was as follows. We first evaluated whether the posterior probability of the evoked response model was greater than that for the null model. This would suggest that time relative to a change point alone is a better predictor of decision dynamics than the average response. If the posterior probability of the evoked response model reliably exceeded the posterior probability of the null model, we then quantified the evidence for alternative models relative to the evoked response model. The sole effect of time relative to a change point was then framed as the new null hypothesis.

We used Bayes Factors to quantify the ratio of evidence for two competing hypotheses. If the ratio is close to 1, then the evidence is equivocal. As the ra-

tio grows more positive, there is greater evidence for the model specified in the numerator, and if the ratio is less than 1, then there is evidence for the model specified in the denominator[78]. Evidence for the null hypothesis is denoted BF_{01} and evidence for the alternative hypothesis is denoted BF_{10}. Because Experiment 2 took a within-subject approach, a separate model was fit for each participant for all proposed models.

To determine whether volatility and conflict affected these peri-change decision dynamics, we modeled changes in decision policy on the drift rate and boundary height surface as a function of λ and p, where λ corresponds to the average period of stability and p corresponds to the mean probability of reward for the optimal choice (see Fig. 2.8 for the full set of models considered). We explored the potential influence of volatility and conflict on the relationship between drift rate and boundary height by examining the posterior probability for each hypothesized model given the set of alternative hypotheses (Model proposals and evaluation; Fig. 2.8A). We found that the evoked response model describing the relationship between shifts in decision parameters and time relative to a change point was more probable than the null model (see Fig. 2.8A).

We also present the evidence for the null model against each alternative model as a Bayes Factor (BF_{01}) (Fig. 2.8B). The 95% confidence interval for the $BF01$ comparing the ratio of evidence for the null model and the evoked response model specifying time-dependent effects of volatility included 1, suggesting inconclusive evidence for either of these models. Likewise, the 95% confidence interval for the BF_{01} comparing the evidence for the null model against the model

specifying change-point-evoked effects of conflict included 1, suggesting no substantive difference between them. Given the equivocal evidence for these two models we excluded them from further comparison with the evoked response model.

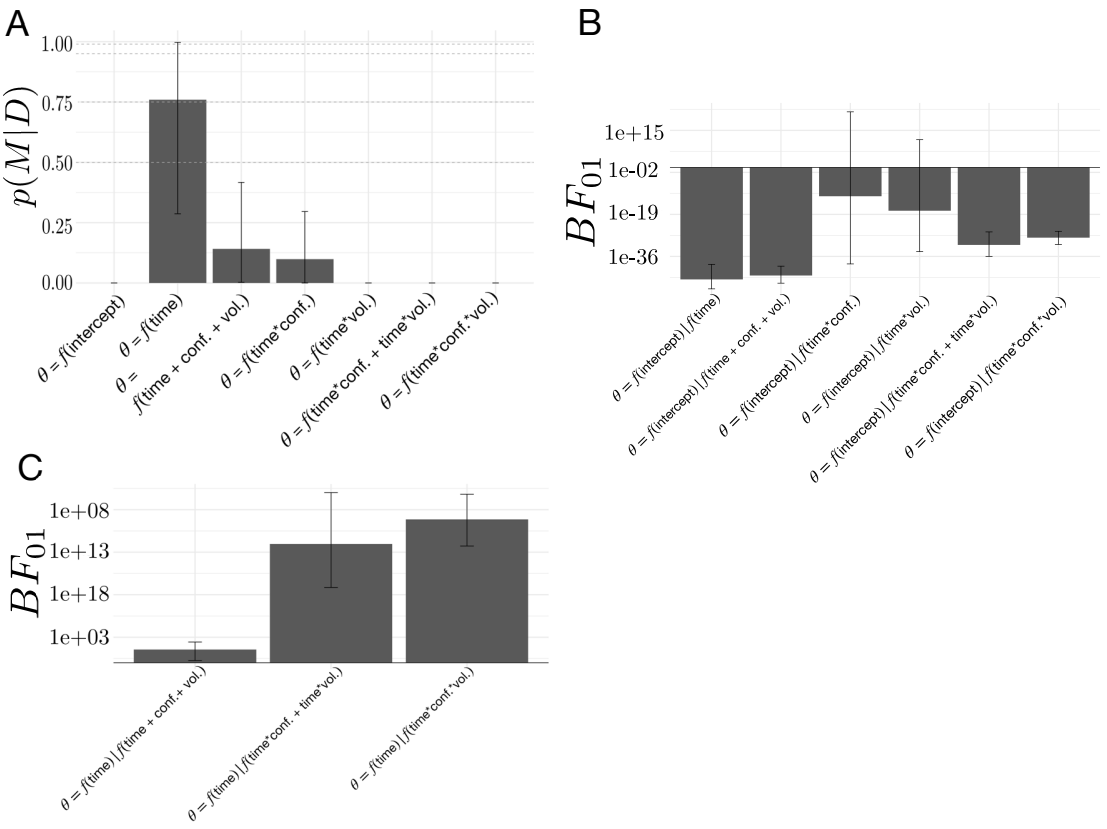


Figure 2.8: Model comparisons for the effect of volatility and conflict on the relationship between drift rate and boundary height. A) The posterior probability for models testing for an effect of volatility and conflict on the angle of shift in a and v, θ. B) The Bayes Factor for the null model relative to the alternative models specifying either an effect of time relative to a change point alone or a conditional effect on this evoked response θ. C) The Bayes Factor for the evoked response model relative to the surviving alternative models specifying a conditional effect on the evoked response, θ. Note that time refers to time relative to the onset of a change point. All models specifying an interaction also include main effects. Dotted horizontal lines refer to grades of evidence[160].

The remainder of the models had substantially negative BF_{01} values (Fig. 2.8B), suggesting that they better fit the data than the null model and allowing them to survive to the next stage of analysis. To evaluate the hypothesis that time alone best accounted for the data, we computed the BF_{01} for the evoked response model against the surviving models from the null model analysis. We find that, for all of the remaining models, the BF_{01} is substantially positive (Fig. 2.8C), indicating that the evoked response model best accounted for the data (posterior probability of evoked response model given the set of models considered: 0.76 ± 0.473; posterior prob. for 3/4 participants > 0.99).

These analyses suggest that the relationship between the rate of evidence accumulation and the boundary height is only related to the change point itself. We find no evidence to suggest that changing the degree of volatility or changing the degree of conflict changes the path of the decision policy following a change point. Thus, the stereotyped response of the decision policy is solely dependent on the presence of a change point rather than either the history of change point frequency or the history of optimal choice ambiguity. Note that while the ideal observer estimates respond to our conditional manipulations of volatility and conflict, the decision dynamics θ we observe do not reflect these effects. This is due to the noisy, imperfect correspondence between the ideal observer signals and a and v. This suggests that adaptation to environmental changes in action-outcome contingencies involves a rapid, coordinated increase in the relationship between the amount of information needed to make a decision and a decrease in the rate of information accumulation, with a stereotyped

return to a stable baseline soon thereafter until another change occurs.

2.2.5 No evidence for locus-coeruleus norepinephrine (LC-NE) system contribution to the decision trajectory

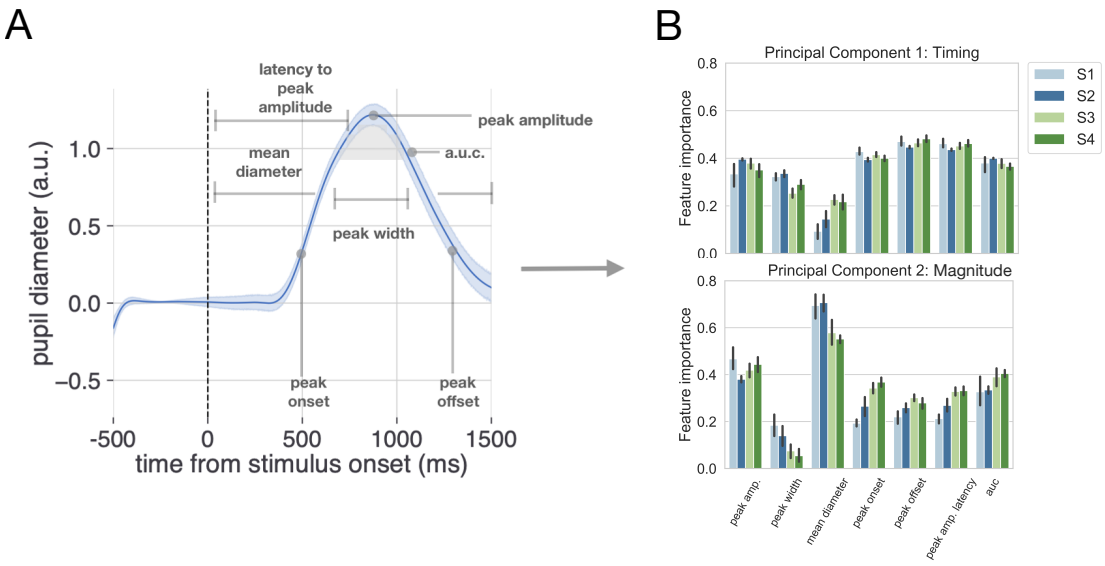


Figure 2.9: Method for analyzing pupil data. A) The evoked pupillary response was characterized according to seven metrics. B) These pupillary features were submitted to a principal component analysis. The contribution of each feature to the variance explained for the first two components is plotted for each subject. Note that we also conducted a supplementary analysis of the task-evoked pupillary response using a more conventional method with similar results.

The LC-NE system is known to modulate exploration states under uncertainty and pupil diameter shows a tight correspondence with LC neuron firing rate[10,125], with changes in pupil diameter indexing the explore-exploit decision state[79]. Similar to the classic Yerkes-Dodson curve relating arousal to performance[175], performance is optimal when tonic LC activity is moderate and phasic LC activity increases following a goal-related stimulus[11], but see[80] for an exception.

Because of this link between LC-NE and the regulation of behavioral variability in response to uncertainty, we expected that LC-NE system responses, as recorded by pupil diameter, would associate with environmental uncertainty and the trajectory through decision policy space following a change in action-contingencies. Specifically, if the LC-NE system were sensitive to a change in the optimal choice, then we should observe a moderate spike in phasic activity following a change in action-outcome contingencies. Note that we do not observe previously established links between exploratory choice behavior and the pupillary response[79,107,156]. We ask the reader to titrate their interpretation of these pupillary data accordingly.

We characterized the evoked pupillary response on each trial in Experiment 2 using seven metrics: the mean of the pupil data over each trial interval, the latency to the peak onset and offset, the latency to peak amplitude, the peak amplitude, and the area under the curve of the pupillary response (see Pupil data preprocessing; Fig. 2.9A). From a computational perspective, reducing the dimensionality of this set of pupillary response metrics expands the set of models we can consider without taxing computational resources in a reasonable amount of time. Further, dimensionality reduction of the pupillary response allows us to capture separable sources of variance relating to timing and amplitude effects without restricting the data to a smaller set of metrics and possibly discarding information (e.g. timing effects may not be constrained to peak latency or onset latency; amplitude effects may not be constrained to peak dilation amplitude). Therefore, we submitted these metrics to principal component analysis to re-

duce their dimensionality while capturing maximum variance.

Evoked response characterization and principal component analyses were conducted for each session and for each subject in Experiment 2. The 95% CI for the number of principal components needed to explain 95% of the variance in the data was calculated over subjects and sessions to determine the number of principal components to keep for further analysis. To aid in interpreting subsequent analysis using the selected principal components, the feature importance of each pupil metric was calculated for each principal component and aggregated across subjects as a mean and bootstrapped 95% CI (Fig. 2.9). We found that the first two principal components explained 95% of the variance in the pupillary data. Peak onset, peak offset, and latency to peak amplitude had the greatest feature importance for the first principal component (Fig. 2.9B, upper panel). Mean pupil diameter and peak amplitude had the greatest feature importance for the second principal component (Fig. 2.9B, lower panel). Thus, for interpretability, we refer to the first and second principal components as timing and magnitude components, respectively (Fig. 2.9B). Note that we also conduct this analysis using more conventional methods of pupillary analysis and continue to observe a null effect (see Pupil data preprocessing for details).

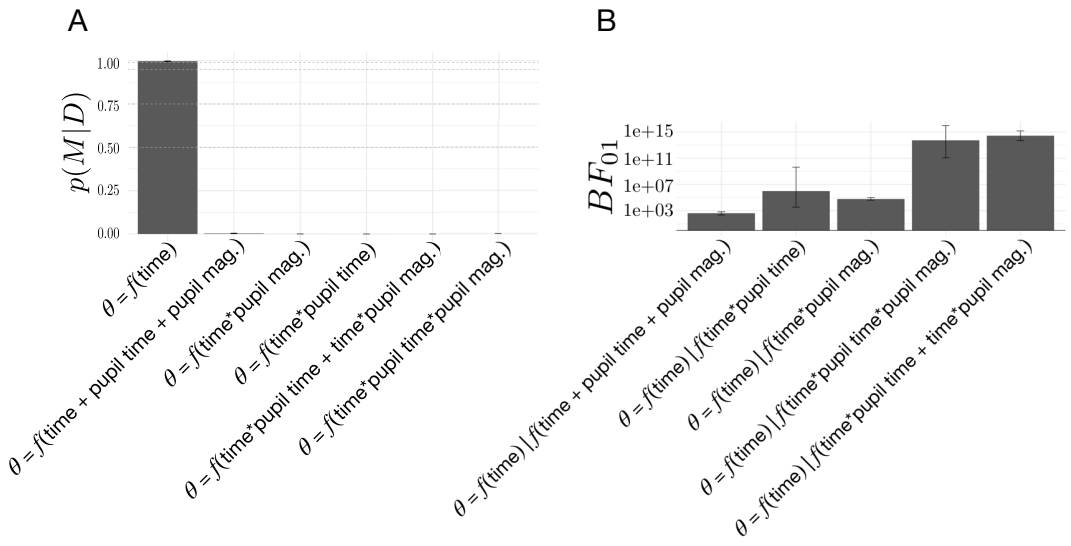


Figure 2.10: Model comparisons for the effect of change-point-evoked pupillary dynamics on the relationship between drift rate and boundary height (θ). A) The posterior probability for models testing for an effect of pupillary dynamics on θ. B) The Bayes Factor for the evoked response model relative to the alternative models specifying an effect of pupillary dynamics on the evoked response, θ. Note that time refers to time relative to the onset of a change point. All models specifying an interaction also include main effects.

To test for the possibility that fluctuations in norepinephrine covaried with changes in the drift-rate and the boundary height, we evaluated a set of models exploring the relationship between the timing and magnitude components of the change-point-evoked pupillary response and shifts in θ. As in our previous model comparison (Fig. 2.8; see Environmental instability prompts a stereotyped decision trajectory), we found that the model describing the relationship between decision policy shift and time relative to a change point had the highest posterior probability given the set of models considered (Fig. 2.10A). To further evaluate the extent of the evidence for the evoked response hypothesis, we present the evidence for the evoked response model against the original model

set as BF_{01} (Fig. 2.10B). We find unambiguous evidence in favor of the evoked response model relative to the models specifying the modulation of θ via the timing and magnitude features of the change-point-evoked pupillary response (posterior probability of time-null model given the set of models considered: 0.997 ± 0.002), with substantially positive BF_{01} values. We find no evidence that the pupillary response associates with the dynamics of the decision policy changes in response to a change in action-outcome contingencies.

2.3 Discussion

We investigated how decision policies change when the rules of the environment change. In two separate experiments, we characterized how decision processes adapted in response to a change in action-outcome contingencies as a trajectory through the space of possible types of exploratory and exploitative decision policies. Our findings highlight how, in the context of two choice paradigms, when faced with a possible change in outcomes, humans rapidly shift to a slow exploratory strategy by reducing the drift rate and, sometimes, increasing the boundary height in a stereotyped manner. Using pupillary data, we were unable to detect a relationship between the LC-NE system and the dynamics of adaptive decision policies in unstable environments. Our findings show how the underlying decision algorithm adapts to different forms of uncertainty.

Exploration and exploitation states are not discrete, but exist along a continuum[2]. Instead of switching between binary states, humans manage environmental instability by adjusting the greediness of their decision policies[133,124,50,170,117,118,168].

Depending on the relative configuration of parameters in the accumulation to bound process, this adjustment can manifest as either speeded or slowed decisions (Fig. 2.1E)[6,126]. Our results suggest that, in the context of volatile two-choice decisions, humans adopt a mechanism that simultaneously changes the rate of evidence accumulation and, sometimes, the threshold of evidence needed to trigger a decision, so as to adapt to an environmental change (Fig. 2.6A). As soon as a shift in action outcomes is suspected, an internal estimate of change point probability increases and an estimate of the belief in the value of the optimal target plummets (Fig. 2.7A). The rapid increase in change point probability causes a rapid *rise* in the boundary height on the subsequent trial, thereby increasing the criterion for selecting a new action and allowing variability in the accumulation process to have a greater influence on choice (Fig. 2.7B), although this latter effect is inconsistent across experiments. These changes lead to slow exploratory decisions that facilitate discovery of the new optimal action and result in a quick recovery of the original threshold value over the course of a few trials. In parallel, the rate of evidence accumulation for the optimal choice decreases, with an immediate drop that gradually returns to its asymptotic value as the belief in the value of the optimal choice stabilizes. These results show that when a learner confronts a change point, the decision policy becomes more exploratory by simultaneously increasing the amount of evidence needed to make a decision *and* slowing the integration of evidence over time. Together, these decision dynamics form a mechanism for gathering information in an unbiased manner that slows the decision at the decision process level but responds

quickly relative to a suspected change in trial time.

Critically, our finding that underlying decision policies can reconfigure multiple underlying decision parameters closely parallels recent work in the domain of information-seeking. Information seeking has been decomposed into random and directed components[170]. Random exploration refers to inherent behavioral variability that leads us to explore other options, while directed exploration refers to the volitional pursuit of new information. Feng and colleagues recently found that random exploration is driven by changes in the drift rate and the boundary height, with drift rate changes dominating the policy shift[50]. When environmental conditions encouraged exploration, the drift rate slowed, reducing the signal-to-noise ratio of the reward representation. This finding clearly aligns with our current observations showing that the drift rate sharply decreases in response to a change point and that this change in drift rate dominates the reconfiguration of decision processes, though our experiments were not designed to isolate the directed and random elements of exploration.

Our results are also broadly consistent with a growing body of research converging on the idea that decision policies are not static, but sensitive to changes in environmental dynamics[45,154]. Previous work by our lab[45] has shown how, during a modified reactive inhibitory control task, different feedback signals target different parts of the accumulation-to-bound process. Specifically, errors in response timing drove rapid changes in the drift rate on subsequent trials, while selection errors (i.e., making a response on trials where the response should be inhibited) changed the boundary height. Further, there is new evidence that

the drift rate adapts on the basis of previous choices, independent of the feedback given for those choices. Urai and colleagues have convincingly demonstrated that choice history signals sculpt the dynamics of the accumulation process by biasing the rate of evidence accumulation[154]. Our current findings and these previous observations[119,127] all highlight how sensitive the parameters of accumulation-to-bound processes are to immediate experience.

Previous literature has shown a conflict-induced spike in reaction time (e.g.[77]). However, our complex reaction time results depart from this. One reason for this departure may relate to the demands of the task we are asking participants to perform. While increased cognitive demand should increase reaction times across conditions, we observe a linear decrease in reaction time as a function of volatility when conflict is highest, and we also see that a net increase in conflict decreases reaction times (Fig. 2.3D). We suspect that the presence of both conflict and volatility blurs the distinction between these two sources of uncertainty, especially under high volatility and high conflict conditions. We also see this effect in our formulation of change point probability (CPP), with a bias to overestimate CPP when conflict is high (Fig. 2.4D). It is possible that participants also exhibit this bias to overestimate volatility when conflict is high, which could muddle the effect of conflict on reaction times. Future research should explore the interaction of change point and conflict estimation on the speed-accuracy trade-off.

We hypothesized that any shift in decision policy in response to a change in action-outcome contingencies would be linked to changes in phasic responses

of the LC-NE pathways[10]. However, we failed to find any evidence of this link using pupilary responses as a proxy of LC-NE dynamics. It should be noted, however, that our experimental design cannot distinguish between pupillary dynamics driven by other catecholamines, such as dopamine, and those dynamics driven by LC-NE system[142,97,64,65], Thus it is possible that the LC-NE system may still be playing a role in shift of decision policies, and the pupil responses we collected were insensitive to the underlying dynamics. Nonetheless, this null association suggests that an alternative neural mechanism drives the adaptive changes that we observed behaviorally.

One possible alternative mechanism for resetting decision policies is is dopaminergic changes to the cortico-basal ganglia-thalamic (CBGT) pathways, or "loops". Both recent experimental[174,43] and theoretical[24,32,163] studies have pointed to the CBGT loops as being a crucial pathway for accumulating evidence during decision making, with the wiring architecture of these pathways ideal for implementing the sequential probability ratio test[23,22], the statistically optimal algorithm for evidence accumulation decisions and the basis for the DDM itself[126]. Further, multiple lines of theoretical work have suggested that, within the CBGT pathways, the difference in direct pathway activity between action channels covaries with the rate of evidence accumulation for individual decisions[100,16,46,130], while the indirect pathways are linked to control of the boundary height[163,73,22,127]. This suggests that changes in the direct and indirect pathways, both within and between representations of different actions, may regulate shifts in decision policies.

Critically, the CBGT pathways are a target of the dopaminergic signaling that drives reinforcement learning[135], suggesting that changes in relative action-value should drive trial-by-trial changes in the drift rate. Indeed, previous work relating dopaminergic circuitry to decision policy adaptation suggests that dopamine may play a critical role in modulating decision policies. Dopamine has substantial links to exploration[81] and recent pharmacological evidence suggests a role for dopaminergic regulation of exploration in humans[35]. More explicitly, both directed and random exploration have been linked to variations in genes that affect dopamine levels in prefrontal cortex and striatum, respectively[64]. Physiologically, previous work has found that a dopamine-controlled spike-timing dependent plasticity rule alters the ratio of direct to indirect pathway efficacy in a simulated corticostriatal network[159], with overall indirect pathway activity (i.e., pre-decision firing rates) linked to the modulation of the boundary height in a DDM and the difference in direct pathway activation across action channels associating with changes in the drift rate[46,130]. Moreover, recent optogenetic work in mice suggests that activating the subthalamic nucleus, a key node in the indirect pathway, not only halts the motoric response but also interrupts cognitive processes related to action selection[74]. Our current observations, combined with this previous work, suggests that the decision policy reconfiguration that we observe may associate with similar underlying corticostriatal dynamics, with belief-driven changes to drift rate varying with the difference in direct pathway firing rates across action channels[46], and change-point-probability-driven changes to the boundary height varying with overall indirect pathway activ-

ity[46,159]. Future physiological studies should focus on validating this predicted relationship between decision policy reconfiguration and CBGT pathways.

The current study raises many more questions about the dynamics of adaptive decision policies than it answers. For example, we only sparsely sampled the space of possible states of value conflict and volatility. Future work would benefit from a more complete sampling of the conflict and volatility space. A psychophysical characterization of how decision states shift in response to varying forms of uncertainty will expose potential non-linear relationships between the decision policy and feedback uncertainty. Moreover, the decisions that we have modeled here are simple two choice decisions, constrained mostly by the normative form of the traditional DDM framework[126]. Scaling the complexity of the task will allow for a more complete assessment of how these relationships change with more complex decisions that better approximate the choices that we make outside the lab. This could be done by moving the cognitive model to frameworks that can fit processes for decisions involving more than two alternatives (e.g.[150]). Finally, because our estimate of the relationship between our ideal observer estimates of uncertainty and human estimates of uncertainty were indirect, this work would benefit from online approximations of ideal observer estimates, as has been done previously[171]. Indeed, there can be substantive individual differences in the detection of of change points[171]. Thus, an approximation of how well the estimates of change point probability from our ideal observer correspond to estimates that human observers hold is needed. This approximation would validate the fidelity of the relationship between the ideal

observer estimates of uncertainty and the decision parameters that we observed.

2.4 Conclusion

Together, our results suggest that when humans are forced to change their mind about the best action to take, the underlying decision policy adapts in a specific way. When a change in action-outcome contingency is suspected, the rate of evidence accumulation decreases and more evidence may briefly be required to commit to a response, allowing variability inherent to the decision process to play a greater role in response selection and resulting in a *slow* exploratory state. As the environment becomes stable, the system gradually adapts to an exploitative state. Importantly, we find no evidence that norepinephrine pathways associate with this response. This suggests that other pathways may be engaged in this adaptive reconfiguration of decision policies. These results reveal the multifaceted underlying decision processes that can adapt action selection policy under multiple forms of environmental uncertainty.

2.5 Methods

2.5.1 Participants

Neurologically healthy adults were recruited from the local university population. All procedures were approved by the Carnegie Mellon University Institutional Review Board (Approval Code: 2018_00000195; Funding: Air Force Research Laboratory, Grant Office ID: 180119). All research participants pro-

vided informed consent to participate in the study and consent to publish any research findings based on their provided data.

Twenty-four participants (19 female, 22 right-handed, 19-31 years old) were recruited for Experiment 1 and paid $20 at the end of four sessions. Four participants (2 female, 4 right-handed, 21-28 years old) were recruited for Experiment 2 and paid $10 for each of nine sessions, in addition to a performance bonus.

Processed data and code are available within a Github repository for this publication. Hypotheses were registered prior to the completion of data collection using the Open Science Framework[56].

2.5.2 Stimuli and Procedure

2.5.2.1 Experiment 1

To begin the task, each participant read the following instructions:

> "You're going on a treasure hunt! You will start with 600 coins in your treasure chest, and you'll be able to pay a coin to open either a purple or an orange box. When you open one of those boxes, you will get a certain number of coins, depending on the color of the box. However, opening the same box will not always give you the same number of coins, and each choice costs one coin. After making your choice, you will receive feedback about how much money you have. Your goal is to make as much money as possible. Press the green button when you're ready to continue. Choose the left box by

pressing the left button with your left index finger and choose the right box by pressing the right button with your right index finger. Note that if you choose too slowly or too quickly, you won't earn any coins. Finally, remember to make your choice based on the color of the box. Press the green button when you're ready to begin the hunt!"

On each trial, participants chose between one of two 'mystery boxes' presented side-by-side on the computer screen (Fig. 2.2A). Participants selected one of the two boxes by pressing either a left button (left box selection) or right button (right box selection) on a button box (Black Box ToolKit USB Response Pad, URP48). Reaction time (RT) was defined as the time elapsed from stimulus presentation to stimulus selection. Reaction time was constrained so that participants had to respond within 100 ms to 1000 ms from stimulus presentation. If participants responded too quickly, the trial was followed by a 5 s pause and they were informed that they were too fast and asked to slow down. If participants responded too slowly, they received a message saying that they were too slow, and were asked to choose quickly on the next trial. In both of these cases, participants did not receive any reward feedback or earn any points, and the trial was repeated so that 600 trials met these reaction time constraints. In order to avoid fatigue, a small break was given midway through each session (break time: 0.70 ± 1.42 m). Participants began each condition with 600 points and lost one point for each incorrect decision.

Feedback was given after each rewarded choice in the form of points drawn

from the normal distribution $N(\mu = 3, \sigma = 1)$ and converted to an integer. If the choice was unrewarded, then participants received 0 points. These points were displayed above the selected mystery box for 0.9 s. To prevent stereotyped responses, the inter-trial interval was sampled from a uniform distribution with a lower limit of 250 ms and an upper limit of 750 ms ($U(250, 750)$). The relative left-right position of each target was pseudorandomized on each trial to prevent incidental learning based on the spatial position of either the mystery box or the responding hand.

To induce decision-conflict, the probability of reward for the optimal target (P) was manipulated across two conditions. We imposed a relatively low probability of reward for the high conflict condition ($P = 0.65$). Conversely, we imposed a relatively high probability of reward for the low conflict condition ($P = 0.85$). For all conditions, the probability of the low-value target was $1 - P$.

Along with these reward manipulations, we also introduced volatility in the action-outcome contingencies. After a prespecified number of trials, the identity of the optimal target switched periodically. The point at which the optimal target switched identities was termed a *change point*. Each period of mean contingency stability was defined as an epoch. Consequently, each session was composed of multiple change points and multiple epochs. Epoch lengths, in trials, were drawn from a Poisson distribution. The lambda parameter was held constant for both high conflict and low conflict conditions ($\lambda = 25$).

To manipulate volatility, epoch lengths were manipulated across two conditions. The high volatility condition drew epoch lengths from a Poisson distri-

bution where $\lambda = 15$ and the low volatility condition drew epoch lengths from a distribution where $\lambda = 35$. In these conditions manipulating volatility, the probability of reward was held constant ($P = 0.75$).

Each participant was tested under four experimental conditions: high conflict, low conflict, high volatility, and low volatility. Each condition was completed in a unique experimental session and each session consisted of 600 trials. Each participant completed the entire experiment over two testing days. To eliminate the effect of timing and its correlates on reward learning[31,109], the order of conditions was counterbalanced across participants.

2.5.2.2 Experiment 2

Experiment 2 used male and female Greebles[59] as selection targets (Fig. 2.2B). Participants were first trained to discriminate between male and female Greebles to prevent errors in perceptual discrimination from interfering with selection on the basis of value. Using a two-alternative forced choice task, participants were presented with a male and female Greeble and asked to select the female, with the male and female Greeble identities resampled on each trial. Participants received binary feedback regarding their selection (correct or incorrect). This criterion task ended after participants reached 95% accuracy (mean number of trials to reach criterion: 31.29, standard deviation over means for subjects: 9.99).

After reaching perceptual discrimination criterion for each session, each participant was tested under nine reinforcement learning conditions composed of

400 trials each, generating 3600 trials per subject in total. Data were collected from four participants in accordance with a replication-based design, with each participant serving as a replication experiment. Participants completed these sessions across three weeks in randomized order. Each trial presented a male and female Greeble[59], with the goal of selecting the sex identity of the Greeble that was most profitable (Fig. 2.2B). Individual Greeble identities were resampled on each trial; thus, the task of the participant was to choose the sex identity rather than the individual identity of the Greeble which was most rewarding. Probabilistic reward feedback was given in the form of points drawn from the normal distribution $N(\mu = 3, \sigma = 1)$ and converted to an integer, as in Experiment 1. These points were displayed at the center of the screen. Participants began with 200 points and lost one point for each incorrect decision. To promote incentive compatibility[76,90], participants earned a cent for every point earned. Reaction time was constrained such that participants were required to respond within 0.1 and 0.75 s from stimulus presentation. If participants responded in $\leq$ .1 s, $\geq$ 0.75 s, or failed to respond altogether, the point total turned red and decreased by 5 points. Each trial lasted 1.5 s and reward feedback for a given trial was displayed from the time of the participant's response to the end of the trial.

To manipulate change point probability, the sex identity of the most rewarding Greeble was switched probabilistically, with a change occurring every 10, 20, or 30 trials, on average. To manipulate the belief in the value of the optimal target, the probability of reward for the optimal target was manipulated,

with P set to 0.65, 0.75, or 0.85. Each session combined one value of P with one level of change point probability, such that all combinations of change point frequency and reward probability were imposed across the nine sessions (Fig. 2.2C). As in Experiment 1, the position of the high-value target was pseudo-randomized on each trial to prevent prepotent response selections on the basis of location.

Throughout the task, the head-stabilized diameter and gaze position of the left pupil were measured with an Eyelink 1000 desktop mount at 1000 Hz. Participants viewed stimuli from within a custom-built booth designed to eliminate the influence of ambient sources of luminance. Because the extent of the pupillary response is known to be highly sensitive to a variety of influences[140], we established the dynamic range of the pupillary response for each session by exposing participants to a sinusoidal variation in luminance prior to the reward-learning task. During the reward-learning task, all stimuli were rendered isoluminant with the background of the display to further prevent luminance-related confounds of the task-evoked pupillary response. To obtain as clean a trial-evoked pupillary response as possible and minimize the overlap of the pupillary response between trials, the inter-trial interval was sampled from a truncated exponential distribution with a minimum of 4 s, a maximum of 16 s, and a rate parameter of 2. The eyetracker was calibrated and the calibration was validated at the beginning of each session. See Pupil data preprocessing for pupil data preprocessing steps.

2.5.3 Models and simulations

2.5.3.1 Q-Learning simulations

A simple, tabular q-learning agent[148] was used to simulate action selection in contexts of varying degrees of conflict and volatility. On each trial, t, the agent chooses which of two actions to take according to the policy

$$\pi_t = \frac{\exp^{\beta * Q_t}}{\Sigma \beta * Q_t}. \tag{2.1}$$

Here β is the inverse temperature parameter, $1/\tau$, reflecting the greediness of the selection policy and Q_t is the estimated state-action value vector on that trial. Higher values of β reflect more exploitative decision policies.

After selection, a binary reward was returned. This was used to update the Q table according using a simple update rule

$$Q_{t+1} = Q_t + \alpha(reward - Q_t), \tag{2.2}$$

where α is the learning rate for the model.

On each simulation an agent was initialized with a specific β value, ranging from 0.1 to 3. On each run the agent completed 500 trials at a specific conflict and volatility level, according to the experimental procedures described in Stimuli and Procedure. The total returned reward was tallied after each run, which was repeated for 200 iterations to provide a stable estimate of return for each agent and condition. The agent was tested on a range of pairwise conflict

($P(optimal) = 0.55 - 0.90$) and volatility ($\lambda = 10 - 100$) conditions.

After all agents were tested on all conditions, the β value for the agent that returned the greatest average reward across runs was identified as the optimal agent for that experimental condition.

2.5.3.2 Drift diffusion model simulations

A normative drift-diffusion model (DDM) process[126] was used to simulate the outcomes of agents with different drift rates and boundary heights. The DDM assumes that evidence is stochastically accumulated as the log-likelihood ratio of evidence for two competing decision outcomes. Evidence is tracked by a single decision variable θ until reaching one of two boundary heights, representing the evidence criterion for committing to a choice. The dynamics of θ is given by.

$$d\theta = vdt + \sigma dW \text{ for } t > tr;$$

$\theta(t \leq tr) = z/a$ (2.3)where v is the mean strength of the evidence and σ is the standard deviation of a white noise process W, representing the degree of noise in the accumulation process. The choice and reaction time (RT) on each trial are determined by the first passage of θ through one of the two decision boundaries $\{a, 0\}$. In this formulation, θ remains fixed at a predefined starting point $z/a \in [0, 1]$ until time tr, resulting in an unbiased evidence accumulation process when $z = a/2$. In perceptual decision tasks, v reflects the signal-to-noise ratio of the stimulus. However, in a value-based decision task, v can be taken

to reflect the difference between Q-values for the left and right actions. Thus, an increase (decrease) in $Q_L - Q_R$ from 0 would correspond to a proportional increase (decrease) in v, leading to more rapid and frequent terminations of θ at the upper (lower) boundary a (0).

Using this DDM framework, we simulated a set of agents with different configurations of a and v. Each agent completed 1500 trials of a "left" (upper bound) or "right" (lower bound) choice task, with $tr = 0.26$ and $z = \frac{a}{2}$. The values for a were sampled between 0.05 and 0.2 in intervals of 0.005. The values for v were sampled from 0 to 0.3 in 0.005 intervals. At the end of each agent run, the probability of selecting the left target, $P(L)$, and the mean RT were recorded.

2.5.3.3 Cognitive model

Our *a priori* hypothesis was that the drift rate (v) and the boundary height (a) should change on a trial-by-trial basis according to two estimates of uncertainty from an ideal observer[27]. We adapted the below ideal observer calculations from a previous study[155] (for the original formulation of this reduced ideal observer model and its derivation, see[112]).

First we assumed that reward feedback drove the belief in the reward associated with an action. We called the belief in the reward attributable to a given action B. This reward belief is learned separately for each action target. Given the chosen target (c) and the unchosen target (u), the belief in the mean reward for the chosen and unchosen targets on the next trial (trial $t + 1$) was calculated as:

$$B_{t+1,c} = B_{t,c} + \alpha_t \delta_t,$$

$\mathrm{B}_{t+1,u} = B_{t,u}(1 - \Omega_t) + \Omega_t E(r)$,

(2.4)where α_t denotes the learning rate, δ_t the prediction error, and Ω_t the change point probability on the current trial t, as discussed below. $E(r)$ refers to the pooled expected value of both targets:

$$E(r) = \frac{\bar{r}_{t_0} + \bar{r}_{t_1}}{2}, \tag{2.5}$$

with $\bar{r}_{t0}, \bar{r}_{t1}$ fixed based on the imposed target reward probabilities.

The prediction error, δ_t, was the difference between the reward obtained for the target chosen and the model belief:

$$\delta_t = r_t - B_{t,c}. \tag{2.6}$$

The signed belief in the reward difference between optimal and suboptimal targets (ΔB) was calculated as the difference in reward value belief between target identities:

$$\Delta B_{t+1} = B_{t,opt} - B_{t,subopt}. \tag{2.7}$$

Model confidence (ϕ) was defined as a function of change point probability (Ω) and the variance of the generative distribution of points (σ_n^2), both of which

formed an estimate of relative uncertainty (RU):

$$RU_t = \frac{\Omega_t\sigma_n^2 + (1-\Omega_t)(1-\phi_t)\sigma_n^2 + \Omega_t(1-\Omega_t)(\delta_t\phi_t)^2}{\Omega_t\sigma_n^2 + (1-\Omega_t)(1-\phi_t)\sigma_n^2 + \Omega_t(1-\Omega_t)(\delta_t\phi_t)^2 + \sigma_n^2}. \tag{2.8}$$

Thus ϕ is calculated as:

$$\phi_{t+1} = 1 - RU_t. \tag{2.9}$$

An estimate of the variance of the reward distribution, σ_t^2, was calculated as:

$$\sigma_t^2 = \sigma_n^2 + \frac{(1-\phi_t)\sigma_n^2}{\phi_t} \tag{2.10}$$

where σ_n is the fixed variance of the generative reward distribution.

The learning rate of the model (α) was determined by the change point probability (Ω) and the model confidence (ϕ). Here, the learning rate was high if either 1) a change in the mean of the distribution of the difference in expected values was likely (Ω is high) or 2) the estimate of the mean was highly imprecise (σ_t^2 was high):

$$\alpha_t = \Omega_t + (1-\Omega_t)(1-\phi_t). \tag{2.11}$$

To model how learners update action-values, we calculated an estimate of how often the same action gave a different reward[155]. This estimate gave our representation of change point probability, Ω. The change point probability approached 1 from below as the probability of a sample coming from a uniform

distribution, relative to a Gaussian distribution, increased:

$$\Omega_t = \frac{U(r_t)H}{U(r_t)H + N(r_t|B_{\Delta_t}, \sigma_t^2)(1-H)}. \tag{2.12}$$

In equation (2.12), H refers to the hazard rate, or the global probability of a change point over trials:

$$H = \frac{n_{cp}}{n_{trials}}. \tag{2.13}$$

Our preregistered expectation was that the belief in the value of a given action and an estimate of environmental stability would target different parameters of the DDM model. Specifically, we hypothesized that the belief in the relative reward for the two choices, ΔB, would update the drift rate, v, or the rate of evidence accumulation:

$$v_{t+1} = \hat{\beta}_v \cdot \Delta B_t + v_t \tag{2.14}$$

while the change point probability, Ω, would increase the boundary height, a, or the amount of evidence needed to make a decision:

$$a_{t+1} = a_0 + \hat{\beta}_a \cdot \Omega_t. \tag{2.15}$$

2.5.3.4 Hierarchical drift diffusion modeling

First, to identify which decision parameters were sensitive to the onset of a change point, we estimated the posterior distribution of drift rate (v), boundary height (a), drift criterion (dc), starting point (z), and non-decision time (t) for the trial preceding the change point in and the following three trials using stimulus-coded fitting methods for Experiment 1. We then looked for change-point-evoked effects in these parameters by comparing the overlap of the distributions for each decision parameter for each of these trials. If less than 5% of the mass of the trial-wise posterior distributions for a given decision parameter overlapped, we considered those distributions to exhibit change point sensitivity.

To identify the fits that best accounted for the data, we conducted a model selection process using Deviance Information Criterion (DIC) scores. We compared the set of fitted models (Table 2.1) to an intercept-only regression model ($DIC_i - DIC_{intercept}$). A lower DIC score indicates a model that loses less information. Here, a difference of ≤ 2 points from the lowest-scoring model cannot rule out the higher scoring model; a difference of 3 to 7 points suggests that the higher scoring model has considerably less support; and a difference of 10 points suggests essentially no support for the higher scoring model[141,30].

We used these complementary model "pruning" methods (i.e. distributional overlap and information loss) as an out-of-set filtering method to determine which decision parameters to include for the subsequent HDDM regression analyses in Experiment 2.

The best parameter fits, evaluated as above, were used to plot the decision trajectory (Decision vector representation) and to estimate the change-point-evoked relationship between those winning parameters (Model proposals and evaluation).

For Experiment 2, to assess whether and how much the ideal observer estimates of change point probability (Ω) and the belief in the value of the optimal target (ΔB) updated the rate of evidence accumulation (v) and the amount of evidence needed to make a decision (a), we regressed the change-point-evoked ideal observer estimates onto the decision parameters using hierarchical drift diffusion model (HDDM) regression[166]. These ideal observer estimates of environmental uncertainty served as a more direct and continuous measure of the uncertainty we sought to induce with our experimental conditions (see Fig. 2.4 for how the experimental conditions impacted these estimates). Considering this more direct approach, we pooled change point probability and belief across all conditions and used these values as our predictors of drift rate and boundary height. Responses were accuracy-coded, and the belief in the difference between targets values was transformed to the belief in the value of the optimal target ($\Delta B_{\text{optimal(t)}} = B_{\text{optimal(t)}} - B_{\text{suboptimal(t)}}$). This approach allowed us to estimate trial-by-trial covariation between the ideal observer estimates and the decision parameters relative to the onset of a change point.

For both the HDDM fits for Experiment 1 and the regression analyses for Experiment 2, Markov-chain Monte-Carlo methods were used to sample the posterior distributions of the regression coefficients. Twenty thousand samples were

drawn from the posterior distributions of the coefficients for each model, with 5000 burned samples and a thinning factor of five. We chose this number of samples to optimize the trade-off between computation time and the precision of parameter estimates, and all model parameters converged to stability. This method generates a distributional estimate of the regression coefficients instead of a single best fit.

To test our hypotheses regarding these HDDM regression estimates, we again used the posterior distributions of the regression parameters. To quantify the reliability of each regression coefficient, we computed the probability of the regression coefficient being greater than or less than 0 over the posterior distribution. We considered a regression coefficient to be reliable if the estimated coefficient maintained the same sign over at least 95% of the mass of the posterior distribution.

2.5.4 Analyses

2.5.4.1 General statistical analysis

Statistical analyses and data visualization were conducted using custom scripts written in R (R Foundation for Statistical Computing, version 3.4.3) and Python (Python Software Foundation, version 3.5.5).

To determine how many trials would be needed to detect proposed condition effects, we conducted a power analysis by way of parameter recovery. For this we simulated accuracy and reaction time data using our hypothesized model (Cognitive model) and calculated the generative or "true" mean drift rate and

boundary height parameters across trials. Then we conducted hierarchical parameter estimation given 200, 400, 600, 800, or 1000 simulated trials. The mean squared error of parameter estimates was stable at 600 trials for all decision parameters. Additionally, as a validation measure, we estimated parameters using component models (drift rate alone, boundary height alone) and a combined model (drift rate and boundary height). We found that the Deviance Information Criterion (DIC) scores among competing models were clearly separable at 600 trials, and in favor of the hypothesized model from which we generated the data, as expected (**??**). Based on these results, we used 600 trials per condition for each participant for our first experiment. We chose to recruit 24 participants for this experiment to fully counterbalance the four conditions ($4! = 24$).

Binary accuracy data were submitted to a mixed effects logistic regression analysis with either the degree of conflict (the probability of reward for the optimal target) or the degree of volatility (mean change point frequency) as predictors. The resulting log-likelihood estimates were transformed to likelihood for interpretability. RT data were log-transformed and submitted to a mixed effects linear regression analysis with the same predictors as in the previous analysis. To determine if participants used ideal observer estimates to update their behavior, two more mixed effects regression analyses were performed. Estimates of change point probability and the belief in the value of the optimal target served as predictors of reaction time and accuracy across groups. As before, we used a mixed logistic regression for accuracy data and a mixed linear regression for reaction time data.

Because we adopted a within-subjects design, all regression analyses of behavior modeled the non-independence of the data as constantly correlated data within participants (random intercepts). Unless otherwise specified, we report bootstrapped 95% confidence intervals for behavioral regression estimates. To prevent any bias in the regression estimates emerging from collinearity between predictors and to aid easy interpretation, all predictors for these regressions were mean-centered and standardized prior to analysis. The Satterthwaite approximation was used to estimate p-values for mixed effects models[134,93].

2.5.4.2 Decision vector representation

To concisely capture the change-point-driven response in the relationship between the boundary height and the drift rate over time, we represented the relationship between these two decision variables in vector space. Trial-by-trial estimates of drift rate and boundary height were calculated from the winning HDDM regression equation and z-scored. Then the difference between each sequential set of (a, v) coordinates was calculated to produce a vector length. The arctangent between these subtracted values was computed to yield an angle in radians between sequential decision vectors (Fig. 2.7B).

For Experiment 1, these computations were performed from the trial prior to the onset of the change point to eight trials after the change point. The initial window of nine trials was selected to maximize the overlap of stable data between high and low volatility conditions (see Supp. Fig. **??**). This resulted in a sequence of angles formed between trials -1 and 0 (Δt_1 yielding θ_1), 0 and 1

(Δt_2 yielding θ_2), and so on. To observe the timescale of these dynamics, a circular regression[106] was performed to determine how θ changed as a function of the number of trials after the change point:

$$\theta = \hat{\beta}_0 + \hat{\beta}_{\Delta_t} + ...\hat{\beta}_{\Delta_{t8}}.$$

To quantitatively assess the number of trials needed for θ to stabilize, we calculated the probability that the posterior distributions of the regression estimates (Supp. Fig. **??**) for sequential pairs of trials had equal means ($\theta_{\Delta_t} = \theta_{\Delta_{t+1}}$). This result (Supp. Fig. **??**) provided an out-of-set constraint on the timescale of the decision response to consider for analogous analyses in Experiment 2.

Experiment 2 used the stability convergence analysis from Experiment 1 to guide the timescale of further circular analyses and, thus, placed a constraint on the complexity of the models proposed (Model proposals and evaluation). Because Experiment 2 took a replication-based approach, a separate model was fit for each participant for all proposed models. We report the mean and 95% CI of the posterior distributions of regression parameter estimates and the mean and standard deviation of estimates across subjects.

The circular regression analyses used Markov-chain Monte-Carlo (MCMC) methods to sample the posterior distributions of the regression coefficients. For both experiments, 10,000 effective samples were drawn from the posterior distributions of the coefficients for each model[89]. Traces were plotted against MCMC iteration for a visual assessment of equilibrium, the autocorrelation function was

calculated to verify independence of MCMC steps, trace distributions were visually evaluated for normality, and point estimates of the mean value were verified to be contained within the 95% credible interval of the posterior distribution for the estimated coefficients.

2.5.4.3 Pupil data preprocessing

Pupil diameter data were segmented to capture the interval from 500 ms prior to trial onset to the end of the 1500 ms trial, for a total of 2000 ms of data per trial. While the latency in the phasic component of the task-evoked pupillary response ranges from 100-200 ms on average[17], suggesting that our segmentation should end 200 ms after the trial ending, participants tended to blink after the offset of the stimulus and during the intertrial interval (see Supp. Fig. **??** for a representative sample of blink timing). Because of this, we ended the analysis window with the offset of the stimulus. Following segmentation, pupil diameter samples marked as blinks by the Eyelink 1000 default blink detection algorithm and zero- or negative-valued samples were replaced by linearly interpolating between adjacent valid samples. Pupil diameter samples with values exceeding three standard deviations of the mean value for that session were likewise removed and interpolated. Interpolated data were bandpass filtered using a .01 to 5 Hz second-order Butterworth filter. Median pupil diameter calculated over the 500 ms prior to the onset of the stimulus was subtracted from the trial data. Finally, processed data were z-scored by session.

For each trial interval, we characterized the evoked response as the mean of

the pupil data over that interval, the latency to peak onset and offset, the latency to peak amplitude, the peak amplitude, and the area under the curve of the phasic pupillary response (Fig. 2.9A). We then submitted these metrics to principal component analysis to reduce their dimensionality while capturing maximum variance. Evoked response characterization and principal component analysis were conducted for each session and for each subject.

The 95% CI for the number of principal components needed to explain 95% of the variance in the data was calculated over subjects and sessions to determine the number of principal components to keep for further analysis.

To aid in interpreting further analysis using the selected principal components, the feature importance of each pupil metric was calculated for each principal component and aggregated across subjects as a mean and bootstrapped 95% CI (Fig. 2.9B).

Note that we also conducted a similar analysis using more conventional methods to assess the task-evoked pupillary response and observed another null effect. Specifically, if we take the derivative of the evoked pupillary response with respect to time[129] and then characterize the pupillary response with the above metrics and conduct principal component analysis, we again see no evidence for a relationship between the pupillary response and the decision trajectory. Additionally, we observe no relationship between our experimental manipulations of conflict and volatility and these metrics, or a change-point evoked shift in pre-stimulus pupillary response ([66], Supp. Fig. **??**). As such, we caution the reader to view our pupillary results in light of this lack of replication of pre-established

exploration-driven pupillary responses.

2.5.4.4 Model proposals and evaluation

To assess the hypothesized influences on θ in Experiment 2, we began our model set proposal with a null hypothesis. Our null model estimates decision dynamics as a function of the intercept, or the average of θ:

$$\theta = \beta_0.$$

Next, we estimated decision dynamics solely as a function of time relative to a change point, with the timescale of consideration determined by the results of the stability convergence analysis from Experiment 1. We call this the evoked response model:

$$\theta = \beta_0 + \beta_{\Delta_t} ... \beta_{\Delta_{tn}}.$$

We first evaluated whether the posterior probability of the evoked response model given the data was greater than the posterior probability for the absolute null model. If the lower bound of the 95% CI of the posterior probability for the time-null model exceeded the upper bound of the 95% CI for the absolute null model (i.e the posterior probability was greater for the evoked response model and the CIs were non-overlapping), we proceeded to evaluate the evidence for alternative models relative to this evoked response model. We evaluated the statistical reliability of the posterior probabilities using a bootstrapped 95% CI

computed over subjects.

We considered an explicit set of hypotheses regarding the effect of the change-point-evoked pupillary response on boundary height and drift rate dynamics (see Fig. 2.10 for the full set of models considered). The first two principal components of the set of pupil metrics, which we term the timing and magnitude components, respectively, were included in this model set to evaluate the effect of the timing and magnitude of noradrenergic dynamics on the change-point-evoked decision manifold. Under the assumption of a neuromodulatory effect on decision dynamics, these principal components were shifted forward by one trial to match the expected timing of the response to neuromodulation.

To determine whether perturbations of volatility and conflict affected change-point-evoked decision dynamics, we estimated the evoked decision dynamics as a function of λ and p, where λ corresponds to the average length of an epoch and p corresponds to the mean probability of reward for optimal target selection (see Table 2.1 for the full set of models considered).

We used Bayes Factors to quantify the ratio of evidence for competing hypotheses[160]. To estimate whether these models accounted for decision dynamics beyond the effect of time relative to a change point alone, we calculate the Bayes Factor for the evoked response model relative to each candidate model (BF_{01}). Finally, we calculate the posterior probability of the null model given the full set of alternative models[160]. Note that this approach assumes that each model has equal *a priori* plausibility.

Bayes Factor visualizations represent the mean and bootstrapped 95% CI

with 1000 bootstrap iterations.

Chapter 3

Cortico-basal ganglia-thalamic (CBGT) network competition supports decision policy reconfiguration

The following text was adapted from the working paper Bond, Rasero, Madan, Bahuguna, Rubin, and Verstynen 2022.

3.1 Introduction

Even the simplest decision relies on complex processing of both external (e.g., sensory) and internal (e.g., learned contingencies) information streams. The choice between two actions is continually updated based on incoming sensory signals at a given accumulation rate until sufficient evidence is reached to trigger one action over the other[67,126]. Importantly, these parameters of information accumulation are highly plastic, adjusting to both the reliability of sensory signals[112,172,111,19,26] and previous choice history[154,127,119,45,46,99], in order to balance the speed of a given decision with local demands to choose the right action.

We recently showed that when action-outcome contingencies change, forcing a change of mind as to what is the most rewarding action, humans dynamically reduce the rate of information accumulation (drift-rate, v, in a normative drift diffusion model, DDM[126]) and, somewhat less reliably, increase the threshold of evidence needed to trigger an action (boundary height, a)[26]. This pushes the decision policy into a slow, exploratory state that allows for feedback learning to push the system to the new best action-outcome contingency, which then leads to a faster drift rate and stable boundary height (see also[45]).

Here we explore the underlying implementation mechanisms that drive changes in underlying decision parameters, as agents have to change their mind in the face of new environmental contingencies. We start with a set of theoretical experiments with the neural systems thought to influence the information accumulation process, known as the cortico-basal ganglia-thalamic (CBGT) networks. These experiments, relying on biologically realistic spiking models of CBGT pathways, make specific predictions that both explain previous results[26] and make specific predictions as to how competition between action channels drive changes in the decision policy. We then test these predictions in humans with neuroimaging using an ultra-high sampled, within-subject design. Each participant is tested across several weeks and several thousands of trials, where action-outcome contingencies change on a semi-random basis. Using whole-brain decoding models, we estimate the relative competition between neural populations encoding choice. Then we show how, consistent with our model predictions, this competition informs underlying decision policy dynamics evoked by environmen-

tal change.

3.2 Results

3.2.1 CBGT circuits control decision parameters under uncertainty

Both theoretical[23,25,127,44,46] and experimental[174] evidence suggests that the CBGT circuits play a critical role in the evidence accumulation process. The canonical CBGT circuit (Figure 1A) includes two dissociable control pathways: the direct (facilitation) and indirect (suppression) pathways[5,57]. A critical assumption of the canonical model is that the basal ganglia are organized into multiple action channels[69,24,165,113,14,123,44,45,158], each containing a direct and indirect pathway. While a strict, segregated action channel organization may not accurately reflect the true underlying circuitry, the concept of independent action channels provides conceptual ease when describing the competition between possible actions without changing the key dynamic properties of the underlying computations. Moreover, striatal neurons have been shown to organize into spatiotemporal assemblies which are modulated by task specific features[3,86,15,33,13]. In the canonical model[101,22], activation of the direct pathway, i.e cortical excitation of D1-expressing spiny projection neurons (SPNs) in the striatum, releases GABAergic signals that can suppress activity in the CBGT output nucleus (internal segment of the globus pallidus, GPi, in primates or substantia nigra pars reticulata, SNr, in rodents). This relieves the thalamus from the tonic inhibition that basal ganglia outputs normally provides, allowing the thalamus to facilitate action execution. Conversely, activation of the indirect pathway,

i.e D2-expressing SPNs in the striatum, controls firing in the external segment of the globus pallidus (GPe) and the subthalamic nucleus (STN) such that it strengthens basal ganglia inhibitory output. This suppresses activity of the thalamocortical pathways, reducing the likelihood that an action gets selected in cortex. The topological encoding of actions in the striatum[110,91,1] and the convergence of projections to the GPi/SNr[85,1] suggests that the direct and indirect pathways may compete for control over the output of the basal ganglia, encoding the "evidence" favoring any behavioral decision as the relative activation of the two pathways within the corresponding action channel[16,43]. Critically, this competition between the pathways has been theoretically linked to the rate of information accumulation during decision making[157].

To illustrate how competition between the direct and indirect pathways regulates information processing during decision making, we designed spiking neural network model of the CBGT circuits, shown in Fig. 4.1A, with dopamine-dependent plasticity occurring at the corticostriatal synapses[159,130], and had it perform a probabilistic dynamic 2-arm bandit task with switching reward contingencies ([26]; see Materials & Methods). This task followed the same general structure as the human participants in the current experiment (see next section) and prior work[26]. In brief, the network selected one of two targets, each of which returned a reward according to a specific probability distribution. The relative difference in reward probability (conflict) was held constant at 75%/25% and probability of a switch in the optimal target (volatility) held at 10 (reward contingency was changed every 10 trials). For purposes of this

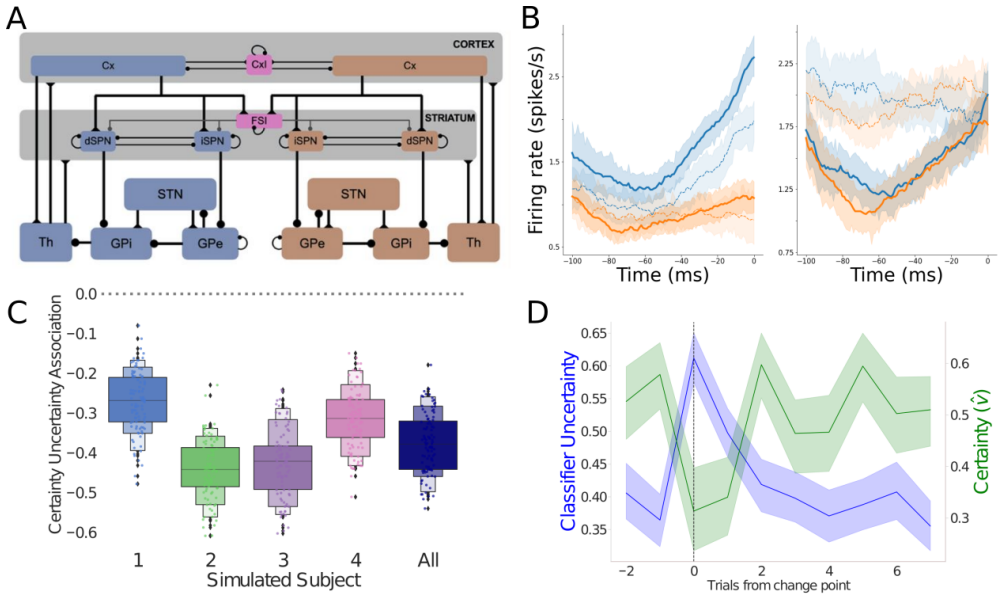


Figure 3.1: Biologically realistic CBGT network performance. (A) Each CBGT nucleus is organized into two action channels (red and blue) except a common population for striatal FSIs (Fast Spiking Interneurons) and cortical interneurons (CxI). CBGT network image adapted from [157]. (B) Average firing rate profiles for D1-SPNs (first column) and D2-SPNs (second column) for trials where left action was chosen, 100ms before the decision time (t=0). The D1-SPNs encoding the "left" action are shown in blue whereas the D1-SPNs encoding the "right" action are shown in orange. The thick solid lines represent the firing rates profiles for fast trials (short RTs) and thin dashed lines represent the firing rates profiles for slow trials (long RTs). The left-dSPNs show a ramping of activity closer to decision time and the slope of this ramp scales with response speed. (C) Drift rates are negatively correlated to decision uncertainty. Simulated subjects represent simulations for different network instances and initial conditions (random seeds). (D) Drift rate and decision uncertainty profiles aligned to the change point. The drift rate drops whereas the decision uncertainty increases as expected at the change point.

study we focus primarily on the effects of switches in optimal targets.

Overall, the network could reliably perform this task, changing its selections in response to a change in action-outcome contingencies effectively (Supp. Fig. 3.5A). Figure 4.1B shows the firing rates of dSPNs and iSPNs in the left action channel, timelocked to selection onset (when thalamic units exceed 30Hz, t=0), for both fast ($<$ 196ms) and slow ($>$ 314.5ms) decisions (see Supp. Fig. 3.5B

for the dynamics of the full network). As expected, the dSPNs show a ramping of activity closer to a left decision for trials where left decision was chosen and the slope of this ramp scales with response speed. In contrast, we see a clearer main effect of response speeds in the iSPNs, which have sustained high firing during slow movements and ramping, rebound firing during fast movements but relatively insensitive to left versus right actions. This is consistent with our previous work showing that differences in direct pathways tracks primarily with choice while differences in indirect pathways modulates overall response speeds[46], the later also supported by experimental studies[176].

In order to capture the parameters of the information accumulation process as the network makes each decision, we modeled the behavior of the CBGT network using a hierarchical version of the drift diffusion model (DDM)[126,167], a canonical formalism for the process of evidence accumulation during decision-making (Fig. 2B). This model returns four key parameters with distinct influences on evidence accumulation, with the drift rate (v) representing the rate of evidence accumulation, the boundary height (a) as the amount of information required to cross the decision threshold, nondecision time (t) as motor-induced delays in the onset of the accumulation process, and starting bias (z) as a bias to begin accumulating evidence for one choice over another (see Methods section).

Consistent with prior observations in humans[26] we found that both v and a were the most pliable parameters across experimental conditions for the network, with the best fitting model showing v being modulated on a trialwise

basis from the estimated difference in reward value (difference in belief of reward value,[112]) and a modulated by the estimation that a change point has occurred[112] (Table 1, left panel). This results in a stereotyped trajectory around a change point (i.e., a change in the optimal target), whereby v immediately plummets and a briefly increases, with a quickly recovering and v slowly returning as reward feedback reinforces the new optimal target (Supp. Figure X). Since prior work has shown that the change in v is more consistently observed than changes in a[26] and since v directly reflects the way information determines the direction of choice, we focus the remainder of our analysis on the control of v.

A central prediction of the opponent pathways models of CBGT circuits is that competition between the direct and indirect pathways will map to the rate and direction of information accumulation during decision-making, i.e., v[44,43,46]. In order to test this, we recorded the average activity of all the 9 subpopulations of the CBGT network during the time between stimulus onset and decision time for all the trials and subjects. The trial-by-trial average activity was used as an input to a L1-penalized logistic principal component regression (LASSO-PCR) classifier to predict the action chosen (left/right) on each trial. The cross-validated accuracies for 4 permutations of our CBGT network (different network instance and initial conditions as a proxy for simulating individual differences) are shown in Figure 4.1C. The trial-wise average activity was able to predict the chosen action with at least 70% accuracy (72-77%) for each initialization of the network, with an an overall accuracy of 74% across

the networks. It should be noted that the activity considered for prediction is the firing during the pre-decision period, which is modulated over trials by reward based plasticity of the corticostriatal weights. A reliable prediction of the trial-wise chosen action indicates that the CBGT network learns the rewarded action within a block as well as when reward contingencies change.

In order to estimate competition between the channels, we took the unthresholded prediction from the LASSO-PCR classifier, $\hat{y}_t$, and calculated its distance from the optimal target (i.e., target with the highest reward probability) on each trial. This provides an estimate of the classifier's certainty of the optimal target, driven by how separable the pre-decision activity is across the network's action channels. Similar degrees of co-activation across the different channels should lead to a greater distance in the classifier's prediction from the optimal target. Thus there should be a negative correlation between classifier uncertainty and v. Across the four permutations of the network, designed to simulate the across-subject variability we expect to observe in the human experiments (see next section), we see that there is indeed a strong negative correlation between these variables (Fig.4.1C). More importantly, when we align both, the classifier uncertainty and v around a change point, we see this negative association is largely driven in response to a change in action outcome contingencies (Fig. 4.1D), consistent with the hypothesis that changes in drift rate are driven by competition between action channels (see also**?**).

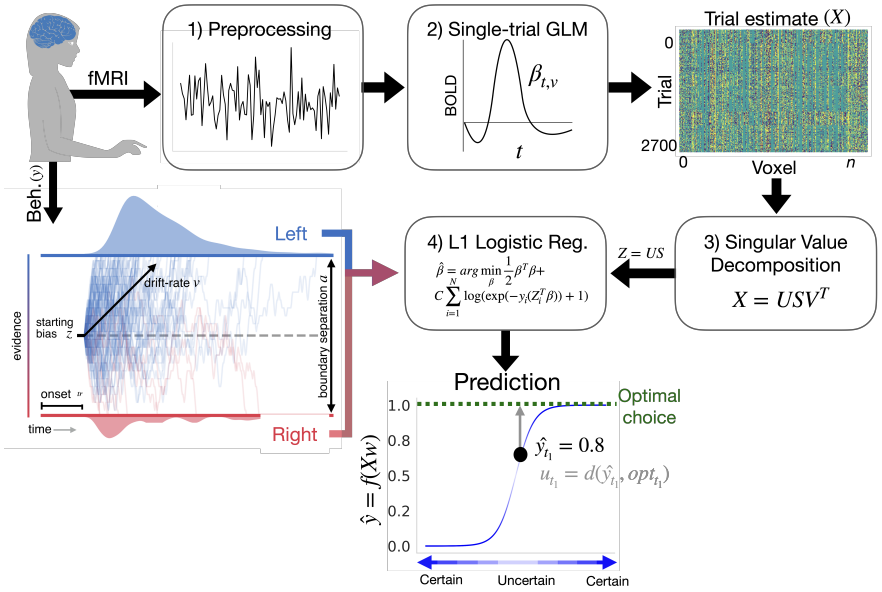


Figure 3.2: Analysis method. Step 1. Preprocessing of fMRI data. Step 2. Single-trial estimates of the hemodynamic response. Step 3. Singular Value Decomposition. Step 4. Logistic regression with an L1 penalty. After crossvalidation, this outputs a predicted response (left or right), here coded as 0 or 1. The further the predicted response from the inflection point of the logistic function, the more certain the prediction. The distance of this predicted response from the optimal choice represents classifier uncertainty for each trial. Here, the predicted probability of a left response $\hat{y_{t1}}$ is 0.2. The distance from the optimal choice on this trial, and, thereby, the classifier uncertainty, u_{t1} is 0.2. Decision parameters were estimated by modeling the joint distribution of reaction times and responses within a drift diffusion framework.

3.2.2 Relative value drives evidence accumulation

Following our previous work[26], we had human participants perform a two-armed bandit task with the same experimental conditions as the CBGT network while we collected trial-evoked BOLD responses using fMRI. Each trial presented a male and female Greeble[59], with the goal of selecting the sex identity of the Greeble that was most rewarding. Participants selected either the left or right Greeble presented on the screen by pressing a button with their left or right hand. To manipulate conflict, rewards were sampled from a Gaussian distribu-

tion for each target. The optimally rewarding target delivered reward with a predetermined probability (P(optimal)) and the suboptimal target gave reward with the inverse probability $(1 - P(\text{optimal}))$ to generate high, moderate, and low conflict $(P(\text{optimal}) = 0.65, 0.75, 0.85)$ conditions. To manipulate volatility, we switched the reward probabilities for the optimal and suboptimal targets according to the rate parameter of a Poisson distribution ( ) to generate high, moderate, and low volatility conditions $(\lambda = 10, 20, 30)$. We tracked internal estimates of action value and environmental volatility using trial-by-trial estimates of two ideal observer parameters, the belief in the value of the optimal choice (ΔB) and change point probability (Ω), respectively (see[26] and Methods for details). Across 45 runs, collected over nine testing sessions, we collected 2700 trials per participant, with the goal of confirming the predictions from our CBGT network simulation in each individual participant.

First, we turn to the effects of our conflict and volatility manipulations on responses. As expected, we found that accuracy increased as contingencies remained stable ($\beta = 0.011, z = 4.919, p = 8.72\text{e}{-07}$; Fig. 4A). Consistent with prior work[26], Ω and ΔB interacted to affect accuracy ($\beta = -0.113, z = -2.271, p = 0.023$) such that the likelihood of a change attenuated the positive relationship between ΔB and accuracy (Supp Fig. X). We also observed a small but statistically reliable increase in reaction times as action-outcome contingencies stabilized ($\beta = 0.001, F = 24.044, p = 9.52\text{e}{-07}$; Fig. 4B). Again, Ω and ΔB interacted to affect reaction times ($\beta = -0.019, F = 15.615, p = 1\text{e}{-04}$). Here, reaction time increased with belief as a change became more likely and re-

action times decreased as belief increased when a change was relatively unlikely (Supp. Fig X).

To address how a change in the environment shifts underlying decision dynamics, we used hierarchical DDM[167] as we did with the network behavior (see Methods for details). Given previous work and the results from our CBGT network model showing that only v and, less reliably, a respond to a shift in the environment[26], we focus on the dynamic regulation of these two parameters on evidence accumulation. Recall that we tracked internal estimates of action value and environmental volatility using two ideal observer parameters, the belief in the value of the optimal choice (ΔB) and change point probability (Ω), as signals that could drive trial-wise changes in DDM parameters. To assess how internal estimates of value and environmental change shift decision dynamics, we estimated single and dual-parameter models mapping ΔB and Ω and drift rate and boundary height (see Table 1). We found that both the dual parameter model mapping ΔB to drift rate and Ω to boundary height and the single-parameter model mapping ΔB to drift rate provided equivocal best fits to the data over human subjects ($\Delta DIC_{\text{null}} = -14.90 \pm 20.58$ and $\Delta DIC_{\text{null}} = -13.80 \pm 16.61$, respectively). All other models failed to provide a better fit than the null model (Table 2).

Again, consistent with prior work[26], we found that the relationship between Ω and the boundary height was unreliable (mean $\beta_{a\sim\Omega} = -0.053 \pm 0.059$; mean $p = 0.448 \pm 0.452$) with a statistically significant decrease in boundary height as Ω increased in only one of four subjects ($\beta_{a\sim\Omega} = -0.122$; $p = 0.002$).

However, drift rate reliably increased with ΔB in three of four subjects (mean $\beta_{v\sim\Delta B} = 0.109 \pm 0.075$; mean $p = 0.082 \pm 0.162$; 3/4 subjects $p < 0.002$). These results suggest that as the belief in the value of the optimal choice approaches the reward value for the optimal choice, the rate of information accumulation increases. We fail to observe a reliable influence of an internal estimate of change on the amount of information required to make a decision.

Altogether, our findings replicate previous work showing that an estimate of value drives the rate of evidence accumulation when a change in the environment is detected. As information about the new optimal choice accrues, the rate of information accumulation increases in parallel, allowing the decision process to assume a more directed path toward the more rewarding option. Using whole brain decoding models, we now evaluate how the relative competition between underlying neural populations encodes the dynamic regulation of evidence accumulation in response to environmental change.

3.2.3 Predicting single trial actions

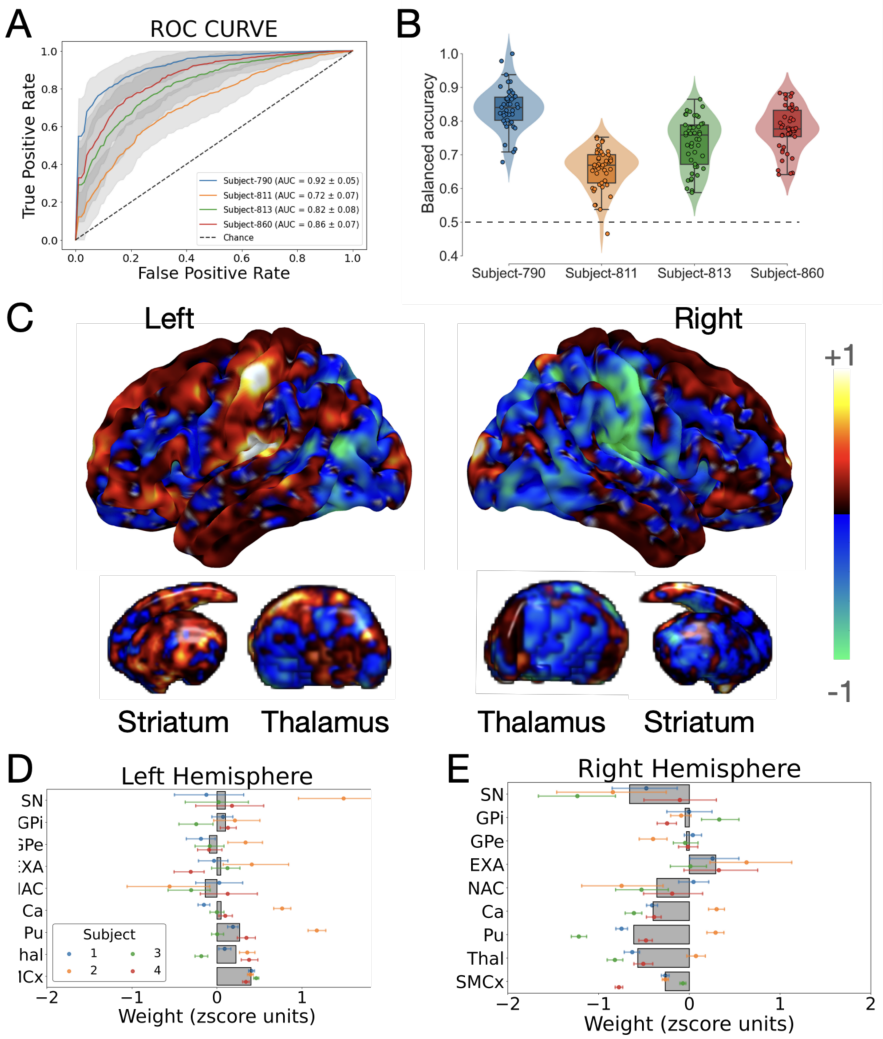


Figure 3.3: Classification performance and feature importance from trial-wise actions. A) The mean cross-validated ROC curve and area under it for classifying single-trial actions for each subject. The black dashed line represents chance performance. B) Balanced accuracy for the classification of trial-wise actions per subject, where each point corresponds to the performance in each cross-validation fold. C) Encoding weight maps in standard space for both hemispheres, averaged across subjects. D) The mean encoding weight and 95% confidence intervals (CI) within regions of interest in the left hemisphere. Points represent individual subjects. Bars display the median across subjects. E) The mean encoding weight and 95% CI within regions of interest in the right hemisphere. SN: Substantia nigra; GPi: Internal segment of globus pallidus; GPe: External segment of globus pallidus; EXA: Extended amygdala; NAC: Nucleus accumbens; Pu: Putamen; Thal: Thalamus; SMCx: Somatomotor cortex.

For each participant, trial-wise responses at every voxel were estimated by means of a general linear model (GLM), with trial modeled as a separate condition in the design input matrix. Therefore, the $\hat{\beta}_{t,v}$ estimated at voxel v reflected the magnitude of the evoked response on trial t. These whole-brain, single-trial responses were then submitted to an L1 penalized principal component logistic regression (Logistic-PCR) to predict left/right response choices. This is the same classifier used to predict the CBGT network responses. In order to reliably generate out-of-sample predictions, the whole predictive modeling was embedded in a leave-one-run-out cross-validation, i.e. each full imaging served as a hold-out test set for each fold of the cross-validation. Generalizability was evaluated as the average across all 45 runs per subject.

The classifier was able to predict single trial responses well above chance (Fig. 4.3A and B), with participant 1 having the best overall performance (AUC = 0.92 ± 0.05, balanced accuracy = 83.88 ± 6.33) and subject 2 the worst (AUC = 0.72 ± 0.07, balanced accuracy = 65.71 ± 5.99), but still well above chance 50% performance. Subjects 3 and 4 showed both comparable, moderate prediction rates (AUC = 0.82 ± 0.08, balanced accuracy = 73.55 ± 7.36; and AUC = 0.86 ± 0.07, balanced accuracy = 77.60 ± 6.51, respectively). Thus, we were able to reliably predict trial-wise responses well above chance for each subject individually.

Once we were able to decode left/right choices from the whole-brain trial-wise responses, we set out to identify which brain regions mainly contributed to such predictions. For this, we fit our classifier to the entire dataset and trans-

formed the obtained weight maps of each subject into encoding patterns, thus ensuring the correct interpretation of the influence of left/right choices in the voxel-wise responses. Fig.4.3C displays the z-scored encoding weight map, average across subjects, where, as expected, contra-lateral brain activations along sensorimotor regions (precentral gyrus) were clearly evoked. Furthermore, quantifying the importance of specific brain regions (Sensorimotor, thalamus and striatum) by averaging the weights within them, one can see that although contributions in the left-hemisphere appeared to be dominated by motor cortex areas (see Fig.4.3D), these were outperformed by subcortical regions in the opposite hemisphere, particularly by the substantia nigra, putamen, caudate and thalamus (See Fig.4.3E). Interestingly, even though the pattern of region-wise contributions was consistent between subjects to a moderate extent, subject 2, whose prediction performance was the lowest, exhibited the most disparate behavior. The importance of the rest of brain regions can be found in Table SXX. Importantly, however, the average patterns across subjects resembled the direction of patterns observed using the same analysis from the CBGT network in contralateral striatum and thalamus (Supp. Fig. 3.5C).

This analysis shows that we could reliably capture trial-wise choices from single-trial BOLD responses on an individual subject basis, replicated across four separate participants. In the next section we used this model to confirm the predictions of the CBGT network.

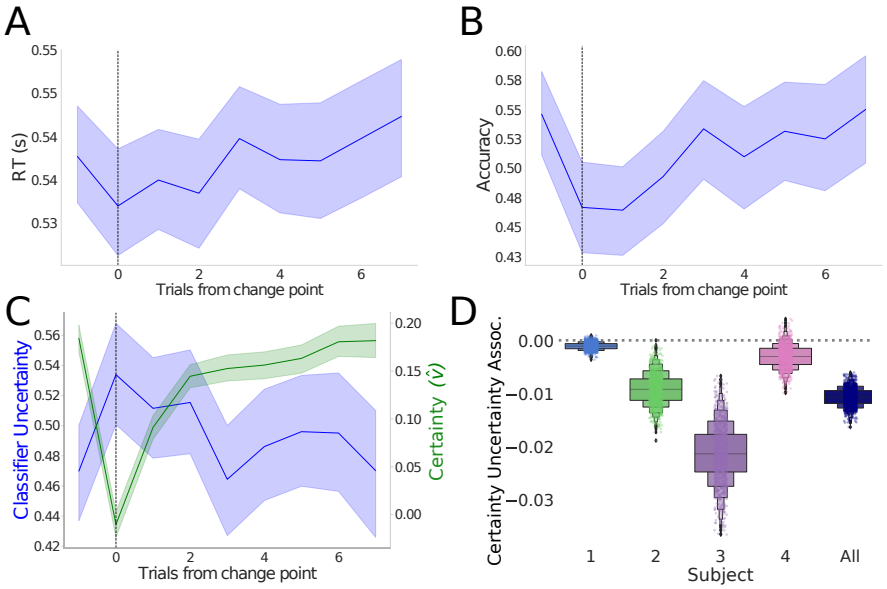


Figure 3.4: Change-point-evoked behavior and certainty. A) Accuracy as the probability of selecting the optimal choice. B) Change-point-evoked reaction times. C) Change-point-evoked classifier uncertainty (blue) and drift rate (v), or certainty (green). D) Bootstrapped distributions of the relationship between decoded classifier uncertainty and certainty (v) by subject and in aggregate.

3.2.4 Competition between action plans drives information accumulation

As a reminder, our core prediction from the CBGT network model is that competition between action channels should correlate with the magnitude of v on each trial. To empirically evaluate this prediction, we examined the link between decision policy parameters and action plan competition decoded from fMRI responses.

We focus on trials around a change point, when action-outcome contingencies switch. The CBGT network model (as well as prior work[26]) predicts that, after the onset of a change-point, v should drop and slowly recover in the direction

of the optimal target and, coincident with this, the uncertainty of the decoder based on human CBGT network responses should spike and reset, leading to an overall negative association between these variables (Fig. 1C).

Consistent with these predictions, we found that v showed a stereotyped response to the onset of a change, with a drastic plummet at the onset of a shift in the optimal choice and a gradual recovery to baseline as action-outcome contingencies remained stable. In parallel, decoder uncertainty increased as the drift rate plummeted (Fig. 4C). Decoder uncertainty was negatively correlated with drift rate in all subjects (Spearman's ρ range: -0.08 to -0.04; p range: 0.00 to 0.043), with an evoked decrease in decoder uncertainty as drift rate increased in 4/4 subjects, to varying degrees (mean bootstrapped β range over subjects: -0.021 to -0.001; t range: -3.996 to -1.326; $p_{S1} = 0.057$, $p_{S2} < 0.000$; $p_{S3} < 0.000$; $p_{S3} = 0.080$, $p_{\text{All}} < 0.000$; Figure 4D).

In sum, we empirically evaluated how competition between neural populations drives dynamic decision policy under changing conditions. Consistent with previous observations, we observe a transition to a slow exploration state at the onset of a suspected shift in action outcome contingencies. This shift is expressed as changes in the rate of information accumulation, or certainty in the decision, with information accumulation slowing at the onset of a change and increasing as the agent learns new action-outcome contingencies. Conversely, we see that uncertainty decoded from neural action plan competition spikes at the onset of a change and gradually returns to baseline as the new properties of the environment are learned. These shifts in decoded uncertainty are negatively associated

with certainty in response to suspected environmental change. This suggests that competition between underlying neural populations may drive a dynamically regulated decision policy to promote learning under changing conditions.

3.3 Discussion

We investigated the implementational mechanisms that may drive decision policy dynamics when the rules of the environment change in humans. Using a high-powered within-subjects design wherein each subject served as an independent replication test, we measured whole-brain hemodynamic responses as participants learned to respond to shifting reward contingencies. First, we found continued support for the way in which decision policy adapts to a suspected change, with a rapid decrease in the rate of evidence accumulation upon detection of a shift, followed by a gradual recovery to baseline rates as the new properties of the environment are learned[26]. Critically, we have empirically validated the theoretical and computational work predicting that competition between neural populations encoding distinct actions modulates decision state[23,25,127,44,46].

While our systems-level approach provides coarse support for the predictions of a biologically realistic CBGT network undergoing similar conditions, the neuronal dynamics of these system-level responses remain to be investigated. As a reminder, the canonical model of the CBGT[22,101] shows that the activation of the direct pathway involves cortical excitation of D1-expressing spiny projection neurons (dSPNs) in the striatum, which then release GABAergic signals that can suppress activity in the CBGT output nucleus, the internal segment of the

globus pallidus, GPi, in primates or substantia nigra pars reticulata, SNr, in rodents. This relieves the thalamus from tonic inhibition, allowing the thalamus to facilitate action execution. Conversely, activation of the indirect pathway, involving D2-expressing SPNs in the striatum, controls firing in the external segment of the globus pallidus (GPe) and the subthalamic nucleus (STN) such that basal ganglia output is inhibited. The topological encoding of actions in the striatum[110,91,1] and the convergence of projections to the GPi/SNr[85,1] suggest that direct and indirect pathways may compete for control over the output of the basal ganglia, encoding the "evidence" favoring any behavioral decision as the relative activation of the two pathways within the corresponding action channel[16,43].

In previous computational studies, the rate of evidence accumulation relied on differences in the simulated ratio of dSPN to iSPN activation in opposing action channels (left or right selection), while the amount of information needed to make a decision relies on overall iSPN activation across action channels[46]. Given that a shift in the rules of the environment most reliably associates with changes in the rate of evidence accumulation[26], the competition we see within the nodes of the CBGT network should correspond to competition between iSPNs and dMSNs in opposing action channels, in contrast with overall iSPN activation over channels. While our current work is suggestive, causal manipulations of dSPN and iSPN competition should test the link between CBGT network competition and decision state.

3.4 Conclusion

The world changes. Therefore, successful adaptation requires flexible decision making. Importantly, the knowledge that the world shifts should be taken into consideration when we weigh the evidence for when to stay with what we know against ex- ploring new options. Altogether, we show that, in both human networks and biologically realistic models of the cortico-basal ganglia-thalamic network, a shift in the environment induces competition between encoded action plans, slowing evidence accumulation to promote adaptive exploration. This work is one step toward understanding the neural computation underlying dynamic decision policy reconfiguration, and thus, flexible decision-making under uncertainty.

3.5 Methods

3.5.1 Participants

Neurologically healthy adults were recruited from the local university population. All procedures were approved by the Carnegie Mellon University Institutional Review Board. All research participants provided informed consent to participate in the study and consent to publish any research findings based on their provided data. Four participants (two female, all right-handed, 29-34 years old) were recruited and paid $30 per session, in addition to a performance bonus and a bonus for completing all nine sessions.

3.5.2 Experimental design

The experiment used male and female Greebles[59] as selection targets (Fig. 1X). Participants were first trained to discriminate between male and female Greebles to prevent errors in perceptual discrimination from interfering with selection on the basis of value. Using a two-alternative forced choice task, participants were presented with a male and female Greeble and asked to select the female, with the male and female Greeble identities resampled on each trial. Participants received binary feedback regarding their selection (correct or incorrect). This criterion task ended after participants reached 95% accuracy. After reaching perceptual discrimination criterion for each session, each participant was tested under nine reinforcement learning conditions composed of 300 trials each, generating 2700 trials per subject in total. Data were collected from four participants in accordance with a replication-based design, with each participant serving as a replication experiment. Participants completed these sessions in randomized order. Each learning trial presented a male and female Greeble[59], with the goal of selecting the sex identity of the Greeble that was most rewarding (Fig. X). Because individual Greeble identities were resampled on each trial, the task of the participant was to choose the sex identity rather than the individual identity of the Greeble which was most rewarding. Probabilistic reward feedback was given in the form of points drawn from the normal distribution $N(\mu = 3, \sigma = 1)$ and converted to an integer. These points were displayed at the center of the screen. For each run, participants began with 60 points and lost one point for each incorrect decision. To promote incentive com-

patibility[76,90], participants earned a cent for every point earned. Reaction time was constrained such that participants were required to respond within 0.1 and 0.75 s from stimulus presentation. If participants responded in $\leq$ .1 s, $\geq$ 0.75 s, or failed to respond altogether, the point total turned red and decreased by 5 points. Each trial lasted 1.5 s and reward feedback for a given trial was displayed from the time of the participant's response to the end of the trial. To manipulate change point probability, the sex identity of the most rewarding Greeble was switched probabilistically, with a change occurring every 10, 20, or 30 trials, on average. To manipulate the belief in the value of the optimal target, the probability of reward for the optimal target was manipulated, with P set to 0.65, 0.75, or 0.85. Each session combined one value of P with one level of volatility, such that all combinations of change point frequency and reward probability were imposed across the nine sessions (Fig. X). Finally, the position of the high-value target was pseudo-randomized on each trial to prevent prepotent response selections on the basis of location.

3.5.3 Behavioral analysis

Statistical analyses and data visualization were conducted using custom scripts written in R (R Foundation for Statistical Computing, version 3.4.3) and Python (Python Software Foundation, version 3.5.5). Binary accuracy data were submitted to a mixed effects logistic regression analysis with either the degree of conflict (the probability of reward for the optimal target) or the degree of volatility (mean change point frequency) as predictors. The resulting log-likelihood

estimates were transformed to likelihood for interpretability. RT data were log-transformed and submitted to a mixed effects linear regression analysis with the same predictors as in the previous analysis. To determine if participants used ideal observer estimates to update their behavior, two more mixed effects regression analyses were performed. Estimates of change point probability and the belief in the value of the optimal target served as predictors of reaction time and accuracy across groups. As before, we used a mixed logistic regression for accuracy data and a mixed linear regression for reaction time data.

3.5.4 Estimating information accumulation using drift diffusion modeling

To assess whether and how much the ideal observer estimates of change point probability (Ω) and the belief in the value of the optimal target (ΔB)[112,26] updated the rate of evidence accumulation (v), we regressed the change-point-evoked ideal observer estimates onto the decision parameters using hierarchical drift diffusion model (HDDM) regression[166]. These ideal observer estimates of environmental uncertainty served as a more direct and continuous measure of the uncertainty we sought to induce with our experimental manipulations. Using this more direct approach, we pooled change point probability and belief across all conditions and used these values as our predictors of drift rate and boundary height. Responses were accuracy-coded, and the belief in the difference between targets values was transformed to the belief in the value of the optimal target ($\Delta B_{\text{optimal(t)}} = B_{\text{optimal(t)}} - B_{\text{suboptimal(t)}}$). This approach allowed

us to estimate trial-by-trial covariation between the ideal observer estimates and the decision parameters.

3.5.5 MRI Data Acquisition

Neurologically healthy human participants (N=4, 2 female) were recruited. Each participant was tested in nine separate imaging sessions using a 3T Siemens Prisma scanner. Session 1 included a set of anatomical and functional localizer sequences (e.g., visual presentation of greeble stimuli with no manual responses, and left vs. right button responses to identify motor networks). Sessions 2-10 collected five functional runs of the dynamic 2-armed bandit task (60 trials per run). Male and female "greebles" served as the visual stimuli for the selection targets[59], with each presented on one side of a central fixation cross. Participants were trained to respond within 1.5 seconds.

To minimize the convolution of the hemodynamic response from trial to trail, inter-trial intervals were sampled according to a truncated exponential distribution with a minimum of 4 s between trials, a maximum of 16 s, and a rate parameter of 2.8 s. To ensure that head position was stabilized and constant over sessions, a CaseForge head case was customized for each participant. The task-evoked hemodynamic response was measured using a high spatial ($2mm^3$ voxels) and high temporal (750ms TR) resolution echo planar imaging approach. This design maximized recovery of single-trial evoked BOLD responses in subcortical areas, as well as cortical areas with higher signal-to-noise ratios. During each functional run, eye-tracking (EyeLink, SR Research Inc.), physiological signals

(ECG, respiration, and pulse-oximetry via the Siemens PMU system) were also collected for tracking attention and for artifact removal.

3.5.6 Preprocessing

fMRI data were preprocessed using the default pipeline of fMRIPrep[48], a standard toolbox for fMRI data preprocessing that provides stability to variations in scan acquisition protocols, a minimal user manipulation, and easily interpretable, comprehensive output results reporting.

3.5.7 Trial-wise responses estimation

By means of a univariate general linear model (GLM) within subject trial-wise responses at the voxel-level were estimated. Specifically, for each fMRI run preprocessed BOLD time series were regressed onto a design matrix, where each task trial corresponded to a different column, and was modeled using a boxcar function convolved with the default hemodynamic response function given in SPM12. Thus, each column in the design matrix estimated the average BOLD activity within each trial. In order to account for head motion, the six realignment parameters (3 rotations, 3 translations) were included as covariates. In addition, a high-pass filter (128 s) was applied to remove low-frequency artifacts. Parameter and error variance were estimated using the RobustWLS toolbox, which adjusts for further artifacts in the data by inversely weighting each observation according to its spatial noise[42].

Finally, estimated trial-wise responses were concatenated across runs and ses-

sions and then stacked across voxels to give a matrix, $\hat{\beta}_{t,v}$, of T (trial estimations) x V (voxels) for each subject.

3.5.8 Single-trial prediction

A machine learning approach was applied to predict left/right greeble choices from the trial-wise responses. First, using the trial-wise hemodynamic responses, we estimated the contrast in neural activation when the participant made a left versus right selection. A L1-constrained principal component logistic regression (Logistic-PCR) was estimated for each subject according to the below procedure. Logistic-PCR with L1 penalty constraint procedure: Dimensionality reduction by a singular value decomposition (SVD) to the input matrix X:

$$X = USV^T \ , \tag{3.1}$$

where the product matrix $Z = US$ represents the principal component scores, i.e. the projected values of X into the principal component space, and V^T an orthogonal matrix whose rows are the principal directions in feature space. Regression of the response binary variable y (Left/Right choice) onto Z, where the estimation of the β coefficients is subject to a L1 penalty term C in the objective function:

$$\hat{\beta} = arg\min_{\beta} \frac{1}{2}\beta^T\beta + C\sum_{i=1}^{N} \log(\exp(-y_i(Z_i^T\beta)) + 1) \ , \tag{3.2}$$

where β and Z include the intercept term, $y_i = \{-1, 1\}$ and N is the number of observations. Projection of the estimated $\hat{\beta}$ coefficients back to the original feature (voxel) space to yield a weight map $\hat{w} = V\hat{\beta}$ used to generate final predictions $\hat{y}$:

$$\hat{y} = \frac{1 - e^{-x \cdot \hat{w}}}{1 + e^{-x \cdot \hat{w}}} \, , \tag{3.3}$$

With x being the vector of voxel-wise responses for a given trial (i.e. a given row in the X matrix).

Here, the competition between left-right neural responses decreases classifier decoding accuracy, as neural activation associated with these actions becomes less separable. Therefore, classifier accuracy serves as a proxy for response competition. To quantify uncertainty from this, we calculated the Euclidean distance of these decoded responses $\hat{y}$ from the statistically optimal choice on a given trial *opt_choice*. This yielded a trial-wise uncertainty metric derived from the decoded competition between neural responses.

$$\hat{U} = d(\hat{y}, opt\_choice). \tag{3.4}$$

The same analytical pipeline was used to calculate single trial responses for simulated data with a difference that trial wise average firing rates of all nuclei from the simulations were used instead of fMRI haemodynamic responses.

3.5.9 Simulations

We simulated neural dynamics and behavior using a biologically realistic cortico-basal ganglia-thalamic (CBGT) network model[45,157]. The network represents the CBGT pathway is composed of 9 neural populations cortical interneurons (CxI), excitatory cortical neurons (Cx), striatal D1/D2-spiny projection neurons (SPNs), striatal fast-spiking interneurons (FSI), the internal (GPi) and external globus pallidus (GPe), the subthalamic nucleus (STN), and the thalamus (Th). All the neuronal populations are segregated into two action channels with the exception of cortical (CxI) and striatal interneurons(FSIs). Each neuron in the population was modeled as an integrate-fire-or-burst-model[163] and a conductance based synapse model was used for NMDA, AMPA and GABA receptors. The neuronal and network parameters (inter-nuclei connectivity and synaptic strengths) were tuned to obtain realistic baseline firing rates for all the nuclei. The details of the model can be referred to in our previous work[157]. Although only one set of tuned network parameters were used for this experiment, the allowed ranges of the parameters have been stated in[157] and the present set of tuned parameters lie well within this range.

Two additional extensions have been added to this CBGT network 1) Spike-timing-dependent plasticity for corticostriatal weights to D1/D2-STR (striatum) that can be modulated by phasic dopamine for reward based learning. 2) an input framework to define common experimental parameters for a 2 arm bandit task (eg.reward probabilities and volatility). The details of the STDP learning have been described in detail in our previous work[159]. As a result of

these extensions the CBGT network can be used to study realistic experimental paradigms with various degrees of decision conflict (reward probabilities) and instability of action values (volatility).

3.5.9.1 Decision threshold

A decision between the two competing actions ("left" and "right") was considered to be made when either of the thalamic subpopulation reached a threshold of 30Hz. This threshold was chosen based on the network dynamics for the chosen parameters with a aim to obtain realistic react times. The maximum time allowed to reach a decision was 1000ms. If none of the thalamic subpopulations reach the threshold of 30Hz, no action was considered to be taken. Such trials were dropped from further analysis. Reaction/decision times were calculated as time from stimulus onset to decision (either subpopulations reach the threshold). The "slow" and "fast" trials were categorized as reaction times ≥ 75 percentile (314.5ms) and reactions time < 50 percentile (196.0ms) respectively of the reaction time distributions. The firing rates of the CBGT nuclei during the reaction times were used for prediction analysis as discussed in section 3.5.8.

3.5.9.2 Corticostriatal weight plasticity

The corticostriatal weights are modified by a dopamine-mediated STDP rule, where the phasic dopamine is modulated by reward prediction error. The internal estimate of the reward is calculated at every trial by a Q-learning algorithm which is subtracted from the reward associated with the experimental paradigm

to yield a trial-by-trial estimate of the reward prediction error. The dopaminergic release is receptor dependent, i.e enables potentiation for D1-SPNs and depression for D2-SPNs. The degree of change in the weights is dependent on an eligibility trace which is proportional to the coincidental pre-synaptic (cortical) and post-synaptic (striatal) firing rates. The STDP rule is described in detail in[159].

3.5.9.3 In silico experimental design

We follow the paradigm of a 2 arm bandit task, where the CBGT network learns to consistently choose the rewarded action until a block change, where the reward contingencies change allowing the CBGT network to show reversal learning. Each session consists of 40 trials with a block change every 10 trials. The reward probabilities represent a conflict of (75%, 25%), eg in a left block, 75% of the left actions are rewarded, whereas 25% of the right actions are rewarded. The inter-trial-interval in network time is fixed to 600ms.

To maximize the similarity between the CBGT network simulations and our human data, we randomly varied the initialization of the network (different network instances with the same connectivity/synaptic conductances) to represent outputs for different simulated subjects.

3.6 Supplementary Figures and Tables

Table 3.1: Simulations

	ΔB	Ω	ΔDIC_{null}	ΔDIC_{best}
I	v	a	-29.85 ±12.76	-4.49 ±5.91
II	a	v	-23.94 ±22.56	-10.40 ±11.22
III	-	v	-6.16 ±4.24	-28.19 ±13.62
IV	v	-	-22.60 ±7.28	-11.74 ±14.80
V	-	a	-7.04 ±11.06	-27.30 ±8.16
VI	a	-	-17.72 ±21.49	-16.62 ±11.88
VII	-	-	0.00 ±0.00	-34.34 ±15.97

Table 3.2: Humans

	ΔB	Ω	ΔDIC_{null}	ΔDIC_{best}
I	v	a	-14.90 ±20.58	-1.52 ±1.04
II	a	v	-0.44 ±1.11	-15.99 ±18.56
III	-	v	-1.47 ±1.30	-14.96 ±18.56
IV	v	-	-13.80 ±16.61	-2.63 ±3.62
V	-	a	-1.03 ±4.46	-15.40 ±15.60
VI	a	-	1.00 ±0.71	-17.42 ±19.52
VII	-	-	0.00 ±0.00	-16.43 ±19.53

Table 3.3: Humans by subject

	Subject	ΔB	Ω	ΔDIC_{null}	ΔDIC_{best}
I	1	v	a	0.61	-2.32
II	1	a	v	0.08	-1.79
III	1	-	v	-1.71	0.00
IV	1	v	-	1.13	-2.84
V	1	-	a	-0.36	-1.35
VI	1	a	-	1.93	-3.64
VII	1	-	-	0.00	-1.71
I	2	v	a	-9.91	-1.73
II	2	a	v	-0.69	-10.95
III	2	-	v	-1.17	-10.47
IV	2	v	-	-11.64	0.00
V	2	-	a	1.89	-13.52
VI	2	a	-	0.46	-12.10
VII	2	-	-	0.00	-11.64
I	3	v	a	-45.08	0.00
II	3	a	v	-1.85	-43.23
III	3	-	v	-3.07	-42.01
IV	3	v	-	-37.41	-7.68
V	3	-	a	-7.53	-37.55
VI	3	a	-	1.16	-46.25
VII	3	-	-	0.00	-45.08
I	4	v	a	-5.23	-2.05
II	4	a	v	0.71	-7.99
III	4	-	v	0.07	-7.35
IV	4	v	-	-7.28	0.00
V	4	-	a	1.90	-9.18
VI	4	a	-	0.43	-7.70
VII	4	-	-	0.00	-7.28

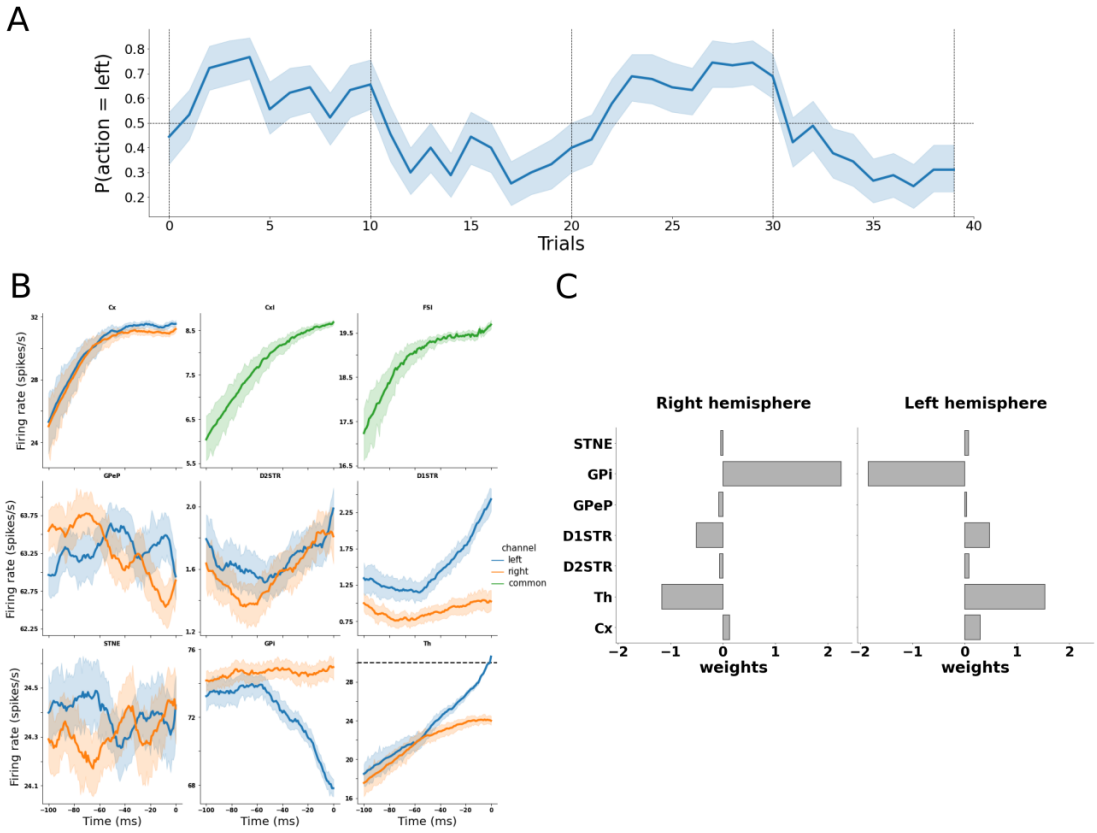


Figure 3.5: CBGT network performance. (A) Choice probability of the CBGT network model in an exemplary session of 40 trial and 4 blocks. The reward contingency (left/right action is rewarded) is changed every 10 trials (marked by vertical dashed lines). The horizontal dashed line represents a chance level (50%) probability to choose left. The trial by trial probability was averaged over many sessions and simulated subjects. The choice probability of choosing left starts at chance level ($\approx$ 50%) when the session begins (trial = 0) but reaches an performance of $\approx$ 70% at the middle of the block. The reward contingency changes every block (every 10 trials), i.e every alternate block (10-20, 30-40) is a block where action right is rewarded. The chocie probability of left action drops during these blocks, because action right is chosen. (B) Firing rate profiles of all the nuclei of the CBGT network for trials where left action was chosen. The decision threshold of 30(spikes/s) is marked by a horizontal dashed line. (C) Encoding weights for CBGT nuclei for predicting the action chosen.

Chapter 4

Decision policy reconfiguration and second-order learning

4.1 Introduction

You can't step into the same river twice. More explicitly, the same intended action may not yield the same reward, and this inherent variability of experience comes not only from the dynamics of the environment we find ourselves in, but also from our own behavior[170] and from internal algorithms, both cognitive[52] and implementational[143,152]. However, the noise properties of living systems can be adaptive. For example, sensory noise in both sensory signals and sensory receptors limits the amount of information available to the central nervous system, allowing filtered computations, acting as a useful constraint on neural computation[49]. Further, behavioral noise can ensure that agents aren't trapped in local minima[169]. More abstractly, signal processing in general can benefit from noise. For example, the concept of stochastic resonance, originally from statistical physics, describes how adding noise to a periodic signal actually

enhances information transfer for weak signals when the input-output system is nonlinear[58], and this idea has been successfully applied to neural systems under the guise of "stochastic facilitation"[98].

In terms of human exploration under uncertainty, recent work has shown that a shift in the rules of the environment changes the underlying decision parameters to adaptively shape choice and response times, promoting adaptive exploration[26,51] (Chapters 2-3). In these situations, the rate of information accumulation slows, resulting in noisiness in the underlying decision process and a relatively unbiased probability of selecting one choice over another. In a reinforcement learning context, this effectively reduces the signal-to-noise ratio of the evidence for the value of each choice. These underlying decision dynamics shape the balance between exploration and exploitation (decision policy) resulting in dynamic adaptation to the shifting demands of our environments.

Less work has been done on how these underlying decision dynamics might shape responses in naturalistic environments with serially dependent choice outcomes. Further, while some work has been done on how these dynamics benefit learning first-order, concrete rules regarding probabilistic stimulus-response pairings[26,51], it isn't clear how these dynamics evolve when learning abstract rules regarding second-order structure in the environment to form second-order decision policy[12].

I investigated how underlying decision dynamics might shape decision policy while learning second-order features of a naturalistic environment with serially dependent decisions. To accomplish this, I asked participants to navigate a spa-

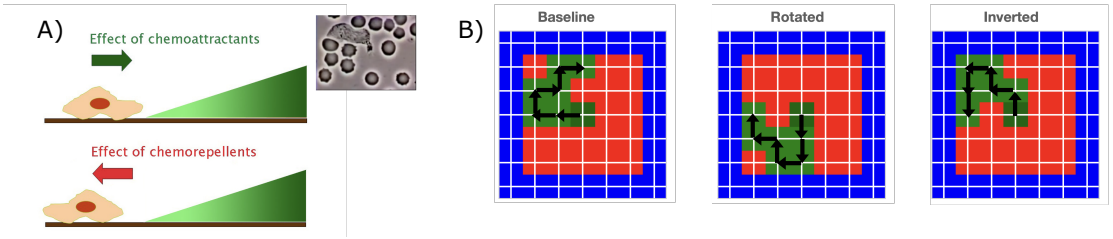


Figure 4.1: Chemotaxis and valeretaxis. A) Chemotaxis, or the movement of an organism in response to a chemical gradient[4]. The landscape of chemoattractants and chemorepellents shapes navigation toward or away from a chemical, respectively. The inset image shows a neutrophil in pursuit of nutrients. Image adapted under the CC 3.0 License. B) Valeretaxis, or action selection in response to a value landscape. The landscape of reward (green cells) and punishment (red cells) shapes action selection. Here, blue cells represent grid walls. Each panel shows an optimal path, annotated by arrows. The left panel shows the baseline reward landscape. The central panel shows a rotation of this baseline reward landscape. The right panel shows an inversion of the optimal path, maintaining a similar degree of complexity as the baseline and rotated paths, but altering path shape.

tial reward landscape to find an optimal path. Much like chemotaxis[4], or movement driven by chemical gradients (Fig. 4.1A), participants were tasked with value-driven spatial navigation, which I term valeretaxis (Fig. 4.1B). To investigate how decision policy evolves while learning and exploiting second-order knowledge of the environment, I ask participants to solve a Baseline path and a Rotated path, which shifts stimulus-response pairings but preserves the second-order feature of path shape. This is followed by an Inverted path that maintains path complexity but perturbs path shape.

4.2 Results

The goal of this experiment was to evaluate how decision policy reconfigures while learning second-order structure. To this end, participants navigated a latent grid space (Fig. 4.1B) under a cover task (Fig. 4.2). First, they were asked

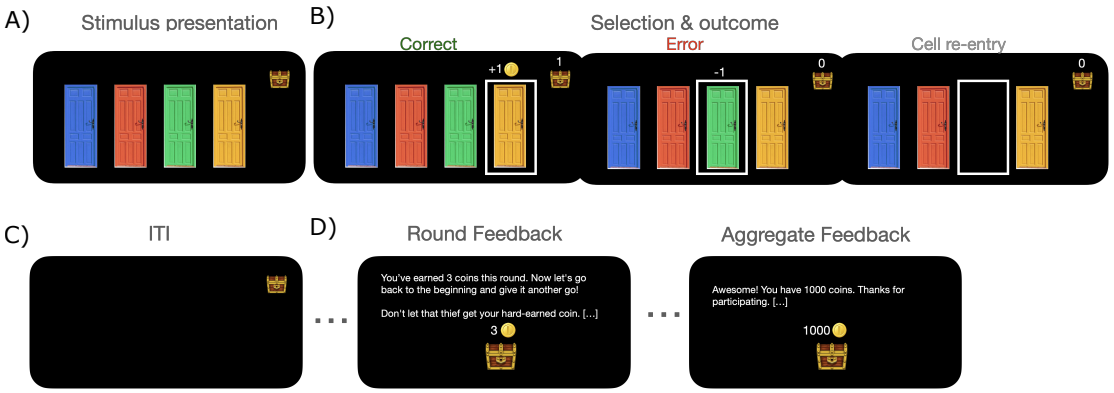


Figure 4.2: Task. A) Participants were presented with a set of four doors that acted as selection arms for spatial navigation, with each door moving the participant left, right, up, or down in latent graph space. Total points for a round were shown above a treasure box on the upper left. B) If participants navigated to a cell on the optimal path, they were rewarded with a coin. Navigating to a cell outside the optimal path was punished with a negative point. Navigating to a cell that had already been visited made the selection arm for that response disappear and the participant received 0 points for that trial. C) Between trials, participants saw a blank screen with a reminder of their point score. D) The left panel shows round-based feedback. Following a round of six choices, the participant was given summative feedback with a reminder of the game reset. The right panel shows aggregate feedback over rounds, displayed at the end of the task.

to find an optimal baseline path (Baseline phase). Then this baseline path was rotated to test if participants could leverage their second-order knowledge of path shape when specific stimulus-response pairings were shifted (Rotation phase). Finally, participants were asked to learn a path of similar complexity but of a different shape (Inversion phase). Participants were given a block of six trials to find the optimal path before they were returned to the center of the grid to begin a new round. Participants were given 68 blocks per phase to learn each map, totaling 1206 trials per subject and 22,914 trials over subjects. Feedback was given trial by trial (Fig. 4.2B) and following each block (Fig. 4.2D), with summary feedback at the end of the session.

Importantly, because this task requires participants to find an optimal *se-*

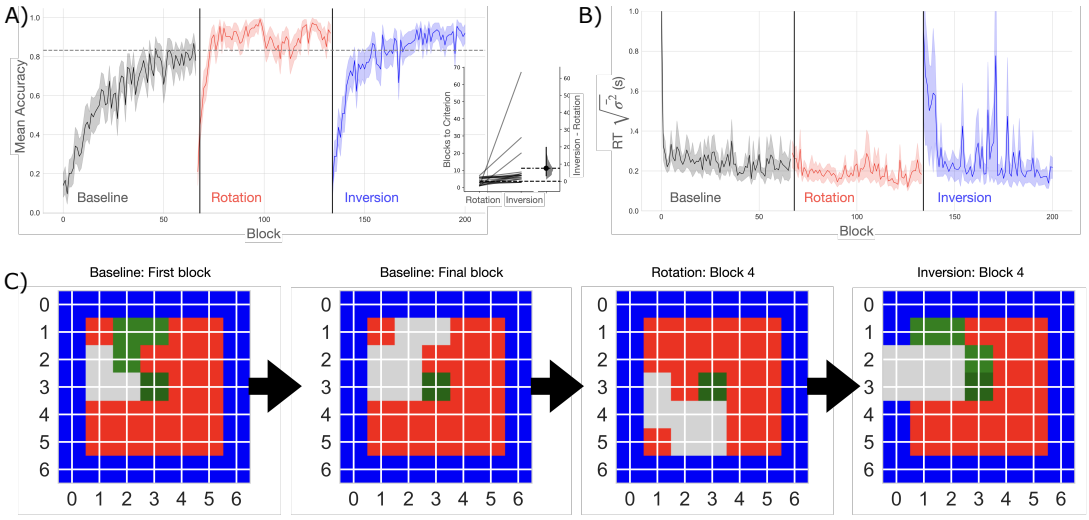


Figure 4.3: Behavior. A) Mean accuracy over blocks. The Baseline phase is shown in black, the Rotation phase is shown in red, and the Inversion phase is shown in blue. The horizontal dashed line marks criterion performance. The inset plot shows a bootstrapped estimate of the pairwise difference in learning rate between the Inverted and Rotated phase, expressed as number of blocks to criterion. Each line represents a single subject. B) A reduced-bias estimate of reaction time variability over blocks by phase. Shaded error shows a bootstrapped estimate of 95% CIs.) Valeretaxis for a single representative subject over time. The optimal path is shown in green and cells selected by the participant are shown in gray. To illustrate initial learning and peak learning in the Baseline phase, the first panel shows path selection in the first block of the Baseline phase, followed by the final Baseline block. The next two plots show early learning in the Rotation and Inversion phases.

quence, the outcomes of their selections depended on previous choices (i.e. outcomes exhibited serial dependence). Because of this serial dependence and the complexity of the demands, this task is also more naturalistic than those employed in Chapters 2-3. See the Methods section for more detail.

4.2.1 Behavior

The logic of this design assumes that, if the learner acquires second-order knowledge about the shape of the path, they should show an accelerated learning rate

when shape (i.e. a second-order feature of the optimal path) is preserved but perturbed in grid space so that specific stimulus-response pairings are no longer relevant. In contrast, if the previously learned path is inverted, altering shape but not complexity, their second-order knowledge will not be helpful and they should show no learning rate advantage.

First, I turn to the observed behavior in terms of accuracy and reaction time. The Baseline map was learned relatively slowly, with participants taking ~ 50 blocks to reach criterion performance ($\sim 84\%$; Fig. 4.3A). Given the difficulty of learning a latent map without navigational cues or an understanding of the selection targets corresponding to directional movements, a protracted familiarization period was to be expected.

Consistent with the hypothesis that second-order knowledge of path shape would benefit learning rate, participants quickly learned the Rotated map, reaching criterion performance in an average of 3.6 ± 2.2 blocks, with a small degree of variability over subjects (Fig. 4.3A, red line). In contrast, the Inversion map was learned relatively slowly, taking 11.2 ± 14.9 blocks to reach criterion, with an elevated degree of variability relative to the Rotation phase (Fig. 4.3A, blue). Indeed, the Rotation map was learned an average of 7.6 blocks faster than the Inverted map (Fig. 4.3A, inset plot).

If the underlying decision policy reconfigures in response to a need to shift stimulus-response pairings, I should also see an expansion and contraction of reaction time (RT) variability following a phase transition. Interestingly, the transition from the Baseline map to the Rotation map shows no reliable change

in RT variability ($\beta = 0.017, t = 1.219, p = 0.224$). On the other hand, the shift from the Rotated map to the Inverted map shows a spike in RT variability during early learning ($\beta = 0.081, t = 2.861, p = 0.004$).

Fig. 4.2C shows qualitative snapshots of the emergence of optimal path-finding in a representative participant, with cell selections shown in gray and the optimal path shown in green. The first two maps show the first block of the Baseline phase and the final Baseline block. Comparing the fourth block of Rotated and Inverted map learning, the discovery of the new optimal path emerges quickly for the path that shares second-order features of the initially learned path. In sum, second-order knowledge appears to benefit learning rate for the Rotation map, which shares second-order features (i.e. path shape). In contrast, the Inverted map, which shares a similar degree of complexity but has different second-order features, shows relatively decelerated learning. Consistent with the idea that underlying decision policy reconfigures when participants need to change their minds regarding the best path to take, reaction time variance expands and contracts when the optimal path shifts from one that shares second-order features with the initially learned map to one that does not. While these behavioral results are suggestive, the next section explicitly tests how underlying decision parameters may reconfigure in response to phase transitions.

4.2.2 Decision dynamics

As a reminder, previous Chapters have shown that underlying decision policy reconfigures upon the detection of a change in the rules of the environment,

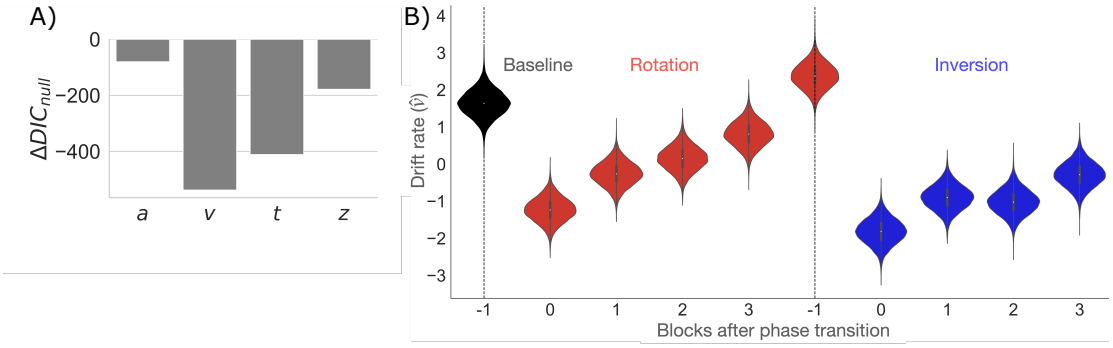


Figure 4.4: Decision dynamics. A) Deviance Information Criterion (DIC) scores for models testing the sensitivity of the four key parameters of the Drift Diffusion Model (DDM), boundary height (a), drift rate (v), non-decision time (t), and starting bias (z). DIC scores are relative to the fit of a null, intercept-only model. B) Block-wise response of drift rate relative to phase transition points, with Baseline estimates in black, Rotation estimates in red, and Inversion estimates in blue. Vertical dashed lines mark blocks prior to phase transitions. A full distribution is shown for each block.

with an initial exploratory phase followed by a gradual shift to exploitation as the new properties of the environment are learned. These decision policy dynamics are largely driven by an initial decrease in the rate of evidence accumulation (v), allowing time for the underlying decision process to diffuse, giving the noisiness of the decision process greater influence over choice to encourage a slow form of exploration. This is followed by a gradual recovery to a relatively elevated baseline, encouraging faster, more exploitative decisions.

This experiment evaluates how underlying decision policy evolves when learning higher order structure in a more naturalistic setting, with choice outcomes dependent on previous choices, as is often the case in the wild. In addition, this experiment tests whether the decision policy dynamics previously observed may benefit learning in environments that share second-order features.

Different underlying decision parameters might shape decision policy in this

context. To examine this possibility, I tested how the four key decision parameters within the Drift Diffusion Modeling (DDM) framework (boundary height, a; the drift rate, v; non-decision time, t; and starting bias, z) responded to phase transitions. As a brief reminder of the meaning of these additional parameters, the boundary height represents the amount of information required to make a decision, non-decision time represents motoric and other non-decision related influences on the decision process, and starting bias shifts the starting point of the decision process toward one choice over another (see the Introduction for an extended description of the DDM framework).

Here, I specifically evaluated how decision parameters responded over the blocks surrounding a phase transition, including the three blocks following the transition. To evaluate the degree of model fit to the observed data, I used the Deviance Information Criterion (DIC), a metric of information loss (see Methods for DDM modeling details). Here, a difference of ≤ 2 points from the lowest-scoring model cannot rule out the higher scoring model; a difference of 3 to 7 points suggests that the higher scoring model has considerably less support; and a difference of 10 points suggests essentially no support for the higher scoring model[141,30]. Scores are described as the difference from a null, intercept-only model.

Consistent with the bulk of my previous work, the model specifying changes in drift rate proximal to a shift in the environment, here as the transitions between Baseline, Rotation, and Inversion phases, clearly showed the best fit to the observational data (Fig. 4.4A). The drift rate model lost significantly less

information than all other models, with the drift model scoring 126.72 points below the second best-fitting model.

Having discovered the drift rate as the best fitting model, I next examined how drift rate changed in response to phase transitions. Generally speaking, drift rate drastically decreased at the phase transition point, with a gradual recovery over the following three blocks and peaking during the final blocks shown (Fig. 4.4B). The transition from the Baseline phase to the Rotation phase decreased drift rate by ~ 2.70 ($p < 0.00$), while the transition from the Rotation phase to the Inversion phase decreased drift rate by ~ 4.19 ($p < 0.000$), showing the greatest change during the shift from Rotation to Inversion. These results show that the onset of phase transition points reduces the rate of information accumulation, with a gradual recovery to faster rates of information accumulation as participants learn the new map. Although these dynamics exist when transitioning between maps that share second-order features, they are more pronounced when participants transition between maps that fail to share these higher-order characteristics.

Altogether, I find initial evidence for decision policy dynamics similar to those discovered in previous Chapters under more naturalistic conditions, with more complex task demands and serially dependent choice outcomes. Intriguingly, these underlying decision dynamics are more pronounced when learners transition between environments without shared second-order features, although these dynamics exist in attenuated form when transitioning between isostructural environments.

4.3 Discussion

I investigated how underlying decision policy evolves while learning second-order environmental structure in a naturalistic setting. My findings provide preliminary evidence for how humans flexibly navigate the explore-exploit continuum in the value-foraging context, or during valeretaxis, the pursuit of value along a spatial reward landscape. These results show the generalizability of the decision algorithm described in Chapters 2-3, where, following a shift in the rules of the environment, humans quickly shift to a slow exploratory state by reducing the rate of evidence accumulation. Further, this work lays the groundwork for understanding how underlying decision policy dynamics are shaped by prior experience.

In the reinforcement learning (RL) context, the concept of cognitive maps encoding the structure of the environment has recently re-emerged as an updated version[105] of the successor representation[38] (SR). SR predicts transitions between states to estimate the optimal trajectory to reward[38,105]. The successor representation (SR) is a reinforcement learning algorithm that builds a predictive map of the environment to summarize the relationship between states separated by multiple state transitions. In Marrian terms, the computational goal of the task put to the participant in this paper is exactly this – the learning (and prediction) of state transitions to find an optimal reward path.

Offline replay, a memory process in which the hippocampal network internally generates patterns of activation representing compressed versions of prior

experience[144], is one neural mechanism thought to support SR[103]. Offline replay has been suggested to combine current experience with previous memories[120] to guide future behavior[104,103]. Importantly, this is not solely a repetition of the past, but a dynamic process sensitive to goal-specification[121] that reverses in response to prediction error[7]. Most relevantly, human and non-human animal studies have shown a role for offline replay in inferring latent environmental structure[104,173], similar to those employed in this paper. Mounting evidence supports the plausibility of SR-related models in both humans and rats as a computational basis for reinforcement learning[41,105]. Future work should capitalize on this research to explore the neural basis of the computations described in this paper, especially exploring how hippocampal dynamics may interact with corticobasal ganglia-thalamic network dynamics previously shown to be relevant to evidence accumulation under uncertainty (Chaper 3).

The experimental design employed in this paper has several limitations. First, note that my experiment does not counterbalance the Rotation and Inversion phases to account for practice and timing effects. However, as designed currently, this means that participants have more path-finding experience for the Inversion phase than the Rotation phase. In this sense, the current design is a stronger test of the metalearning effect the experiment is designed to evaluate. Still, future work should counterbalance these two test phases to minimize fatigue contaminating learning in the final phase.

Second, my experiment does not clearly distinguish between sequence learning and learning the latent graph reward landscape. A high-performing learner

might simply be learning an optimal sequence of button presses rather than underlying graph space, as learning both a purely sequence-based representation and learning a graph representation yields the same behavioral output. To dissociate these possibilities, follow-up experiments will perturb the learner's position on the graph after they demonstrate learning, or intermittently throughout learning. If they truly know the graph space, then they should be able to reorient toward the optimal path. If they simply learned an optimal sequence, then performance should decrement with slower recovery.

4.4 Conclusion

These results expand my previous work to show that, under more naturalistic conditions with serial dependence between choice outcomes, the underlying decision policy maintains a stereotyped response to change. As the reward landscape shifts, evidence accumulation rates decrease, allowing noise in the underlying decision process to influence response selection as a slow form of exploration. As the new properties of the environment are learned, evidence is accumulated at a faster rate to promote exploitative choice.

Further, the evolution of underlying decision policy appears to be influenced by prior experience with environments that share second-order features with the current context.

Altogether, this work shows the generalizability of adaptive decision policy reconfiguration and establishes a basis for further investigating how structural similarities between environments influence underlying decision policy dynamics

to promote generalization.

4.5 Methods

4.5.1 Participants

Twenty neurologically healthy adults (18-35 years old) were recruited from the local university population and paid $10 per session with a performance bonus of $0.01 per point earned. All procedures were approved by the Carnegie Mellon University Institutional Review Board. All research participants provided informed consent to participate in the study and consent to publish any research findings based on their provided data.

4.5.2 Stimuli and Procedure

The goal of this experiment was to test if the decision policy reconfiguration observed in Chapters 2-3 supported exploration in service of learning second order structure of the environment. To assess this, I first asked participants to navigate a grid world with an optimally rewarding path to teach them a higher order structure in terms of path shape (Fig. 4.1B). Following this baseline learning phase, I rotated the path (Rotation phase) and then exposed them to a path of similar complexity but a different shape (Inversion phase) (see Fig. 4.1B, Rotated and Inverted phases).

On each trial, the participant chose between one of four doors of different colors arranged as shown in Fig. 4.2 by pressing one of four buttons on a button box (Black Box ToolKit USB Response Pad, URP48). Each door acted as the

cue for a movement (Up, Down, Left, or Right) in grid space. Critically, participants were naive to the fact that they were navigating a latent graph and they were not informed that their selections corresponded to movements over the latent map. To ensure implicit navigation of the reward landscape, there was a cover task as described in the instructions below:

> "You're going on a treasure hunt! In this hunt, you can choose to open one of four colored doors. Opening one of these doors may reveal a coin you can add to your chest. However, opening the same door will not always give you the same number of coins.
>
> In fact, there's a thief afoot! The thief sometimes steals the coins you already have. But they don't stop there. Other times, the thief even tries to block your access to what might be behind a given door at a certain point in time. When they block you, the door vanishes. So choose carefully!
>
> After making your choice, you will receive feedback about how many coins you earned or lost, with a summary of your earnings after every 6 choices. Your goal is to gather as many coins as possible."

If the participant was rewarded for their choice, they earned one coin and this was displayed above their selection (see Fig. 4.2B). If they navigated outside of the optimal path, one coin was removed from their total. If they hit a wall at the edges of the grid (blue cells in Fig. 4.1B)), the door for that choice disappeared. Because participants were attempting to find an optimal *sequence* of

decisions (i.e. path), if they visited a previously "consumed" cell, they earned 0 points. Feedback was displayed for 0.9 s. To prevent stereotyped responses, the inter-trial interval was sampled from a uniform distribution with a lower limit of 250 ms and an upper limit of 750 ms ($U(250, 750)$).

Reaction time was constrained so that participants had to respond within 100 ms to 1000 ms from stimulus presentation. If participants responded too quickly, the trial was followed by a 5 s pause and they were informed that they were too fast and asked to slow down. If participants responded too slowly, they received a message saying that they were too slow, and were asked to choose quickly on the next trial. In both of these cases, participants did not receive any reward feedback or earn any points.

Participants were given six trials to find the optimal path of six cells. Each round of six trials was a block. After each block, the total point feedback for that round was presented and they were informed that they were starting afresh with a new round. At the beginning of each round, they were returned to the center of the grid. Each participant completed 201 blocks, with 402 trials (67 blocks) in each of the Baseline, Rotation, and Inversion phases. For the purpose of this preliminary experiment, each participant solved the same set of Baseline, Rotated, and Inverted paths. This yielded 1206 trials per subject and 24,120 trials in aggregate.

4.5.3 Analyses

4.5.3.1 Behavior

Behavioral performance was evaluated using reaction time and accuracy. Reaction times were calculated as the interval between stimulus presentation and button press. Accuracy was calculated as the cell-selection overlap with the cells composing the optimal path (discounting repeated cell entries). Criterion-based analyses relied on an estimate of chance accuracy for path selection as $\frac{1}{6}$, or $\sim$ 16%. Participants were considered to have reached criterion when their accuracy reached $\sim$ 84% over a block.

To assess differences in learning rate between phases, I used a paired t-test evaluating the number of blocks to criterion for both the Rotation and Inversion phases. Significance was evaluated by permutation, with 5000 shuffles of the values for the Inversion and Rotation phases. P-values represent the likelihood of observing the effect size if the null hypothesis of zero difference between phases is true. Bootstrapped confidence intervals for the magnitude of the difference between Inversion and Rotation phases are bias-corrected and accelerated, with 5000 resamples.

To minimize estimator bias, reaction time variability for each block was calculated as the square root of the mean variance over subjects. Differences in this estimate of RT variance were assessed using a simple linear regression with the Rotation and Inversion phases as predictors and RT variance as the outcome.

4.5.3.2 Modeling

As in previous Chapters, I evaluated a set of hierarchical Drift Diffusion Models (DDM) using information loss criteria to assess the degree of model fit to the data. I use Deviance Information Criterion (DIC) scores for this evaluation because they are well-suited to measuring model fit in hierarchical models[160]. Instead of relying on a single maximum likelihood estimate, I used Markov Chain Monte Carlo (MCMC) sampling to generate a distribution for each estimate. In total, 9,500 effective samples were drawn from the posterior distributions of the coefficients for each model, with the first 500 samples used as burn-in to ensure stability of fits[89]. Because convergence, or parameter stability, is crucial for interpreting results using this approach, traces were plotted against MCMC iteration for a visual assessment of equilibrium, the autocorrelation function was calculated to verify independence of MCMC steps, trace distributions were visually evaluated for normality, and point estimates of the mean value were verified to be contained within the 95% credible interval of the posterior distribution for the estimated coefficients.

To evaluate the significance of posterior distributions, I calculated an empirical probability for each estimate by summing the number of estimate samples in the same direction as the mean estimate and dividing by the total number of samples. If the sign of the average value was maintained for 95% of the distribution, I considered that parameter significant. Likewise, if less than 5% of two distributions overlapped, I considered them reliably different from one another. This was supplemented by permutation testing, as described above.

To identify the model fits that best accounted for the data, I conducted a model selection process using Deviance Information Criterion (DIC) scores. We compared the set of fitted models to an intercept-only regression model ($DIC_i - DIC_{intercept}$). A lower DIC score indicates a model that loses less information. Here, a difference of ≤ 2 points from the lowest-scoring model cannot rule out the higher scoring model; a difference of 3 to 7 points suggests that the higher scoring model has considerably less support; and a difference of 10 points suggests essentially no support for the higher scoring model[141,30].

Chapter 5

Conclusion

The goal of this dissertation was to investigate how humans flexibly balance the adaptive value of noise (exploration) with the value of acting on what they know (exploitation) to dynamically adapt to changing conditions. I sought to address this question with three aims spanning the algorithmic and implementational levels of explanation. First, using a reinforcement-learning-driven evidence accumulation framework, I tested how decision policy reconfigures in response to a shift in the environment. Then, I tested the replicability of these findings and explored the underlying implementational mechanisms for the dynamic decision policy reconfiguration observed. Finally, using a foraging experiment, I explored how dynamic decision policy reconfiguration may also support learning more abstract, second-order structure in the environment.

Together, this work shows that underlying decision policy evolves in a stereotyped manner in response to change with evidence for the CBGT network competition driving this response. This decision policy reconfiguration also appears to support metalearning. In total, the experiments in this dissertation eluci-

date one mechanism for modulating exploration under uncertainty, and take first steps toward understanding how this mechanism may support second-order learning.

In Chapter 2, I test how underlying decision policy evolves in response to a change in action-outcomes contingencies and evaluate how the locus-coeruleus norepinephrine system may modulate these responses. I find that, when a change in the environment is suspected, evidence is accumulated more slowly over time in order to promote exploration, with a return to baseline rates of evidence accumulation as the new action-outcome contingencies are learned. At times, this response is accompanied by an increase in the amount of evidence required to make a decision, allowing greater time for the decision process to diffuse. This change-evoked decrease in evidence accumulation rate replicates under superficially different task conditions and over three subjects, each as an independent out-of-set test of the effect.

In Chapter 3, I explore how corticobasal-ganglia thalamic network dynamics associate with this adaptive reconfiguration. Altogether, I show that, in both human networks and biologically realistic models of the cortico-basal ganglia-thalamic network, a shift in the environment induces competition between encoded action plans, slowing evidence accumulation to promote adaptive exploration. This work is one step toward understanding the neural computation underlying dynamic decision policy reconfiguration, and thus, flexible decision-making under uncertainty.

Finally, Chapter 4 goes up one layer of abstraction and tests how this de-

cision policy may reconfigure when learning higher order structure. These results expand my previous work to show that, under more naturalistic conditions with serial dependence between choice outcomes, the underlying decision policy maintains a stereotyped response to change. As the reward landscape shifts, evidence accumulation rates decrease, allowing noise in the underlying decision process to influence response selection as a slow form of exploration. As the new properties of the environment are learned, evidence is accumulated at a faster rate to promote exploitative choice. Further, the evolution of underlying decision policy appears to be influenced by prior experience with environments that share second-order features with the current context. This chapter is a preliminary pass at showing the generalizability of adaptive decision policy reconfiguration. Further, it establishes a basis for the future investigation of how structural similarities between environments influence underlying decision policy dynamics to promote generalization.

Altogher, this work supports a growing body of evidence that decision policies are dynamic functions that move along the exploration-exploitation continuum to adapt to changes in environmental dynamics ([45,87,114,154,119,127]. Most importantly, my completed work expands on these observations by characterizing, for the first time, the decision dynamics evoked in response to a suspected shift in the rules of the environment. The work completed has assessed these dynamics at the implementational level for two plausible neuromodulatory systems. More generally, linking the adaptive reconfiguration of the processes underlying a decision to changes in decision policy provides a scaffold to explore the mechanism

driving shifts in action selections in response to environmental change.

References

[1] (2021). The mouse cortico – basal ganglia – thalamic network. *Nature*, 598(October).

[2] Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L., & Platt, M. L. (2017). A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology*, 42(10), 1931–1939.

[3] Adler, A., Finkes, I., Katabi, S., Prut, Y., & Bergman, H. (2013). Encoding by synchronization in the primate striatum. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(11), 4854–66.

[4] Adler, J. (1975). Chemotaxis in bacteria. *Annual review of biochemistry*, 44(1), 341–356.

[5] Albin, R. L., Young, A. B., & Penney, J. B. (1995). The functional anatomy of disorders of the basal ganglia. *Trends in Neurosciences*, 18(2), 63–64.

[6] Alexandrowicz, R. W. (2020). The diffusion model visualizer: an interactive tool to understand the diffusion model parameters. *Psychological research*, 84(4), 1157–1165.

[7] Ambrose, R. E., Pfeiffer, B. E., & Foster, D. J. (2016). Reverse replay of hippocampal place cells is uniquely modulated by changing reward. *Neuron*, 91(5), 1124–1136.

[8] Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, 543(7647), 719–722.

[9] Aston-Jones, G. & Bloom, F. (1981). Activity of norepinephrine-containing locus coeruleus neurons in behaving rats anticipates fluctuations in the sleep-waking cycle. *Journal of Neuroscience*, 1(8), 876–886.

[10] Aston-Jones, G. & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, 28, 403–450.

[11] Aston-Jones, G., Rajkowski, J., & Cohen, J. (1999). Role of locus coeruleus in attention and behavioral flexibility. *Biological psychiatry*, 46(9), 1309–1320.

[12] Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, 66(2), 315–326.

[13] Badreddine, N., Zalcman, G., Appaix, F., Achard, S., Fino, E., Badreddine, N., Zalcman, G., Appaix, F., Becq, G., & Tremblay, N. (2022). Spatiotemporal reorganization of corticostriatal networks encodes motor skill learning. *Cell Reports*, 39.

[14] Bahuguna, J., Weidel, P., & Morrison, A. (2019). Exploring the role of striatal D1 and D2 medium spiny neurons in action selection using a virtual robotic framework. *European Journal of Neuroscience*, 49(6), 737–753.

[15] Barbera, G., Liang, B., Zhang, L., Gerfen, C., Culurciello, E., Chen, R., Li, Y., & Lin, D.-T. (2016). Spatially Compact Neural Clusters in the Dorsal Striatum Encode Locomotion Relevant Information. *Neuron*, 92(1), 202–213.

[16] Bariselli, S., Fobbs, W., Creed, M., & Kravitz, A. (2018). A competitive model for striatal action selection. *Brain research*.

[17] Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2), 276.

[18] Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2), 490–509.

[19] Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9), 1214.

[20] Berlyne, D. E. (1978). Curiosity and learning. *Motivation and emotion*, 2(2), 97–175.

[21] Bland, A. R. et al. (2012). Different varieties of uncertainty in human decision-making. *Frontiers in neuroscience*, 6, 85.

[22] Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends in cognitive sciences*, 11(3), 118–125.

[23] Bogacz, R. & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural computation*, 19(2), 442–477.

[24] Bogacz, R. & Larsen, T. (2011). Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. *Neural computation*, 23(4), 817–851.

[25] Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in neurosciences*, 33(1), 10–16.

[26] Bond, K., Dunovan, K., Porter, A., Rubin, J. E., & Verstynen, T. (2021). Dynamic decision policy reconfiguration under outcome uncertainty. *Elife*, 10, e65540.

[27] Bond, K., Dunovan, K., & Verstynen, T. D. (2018). The influence of volatility and conflict on adaptive decision making.

[28] Bouret, S. & Sara, S. J. (2005). Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. *Trends in neurosciences*, 28(11), 574–582.

[29] Braun, D. A., Mehring, C., & Wolpert, D. M. (2010). Structure learning in action. *Behavioural brain research*, 206(2), 157–165.

[30] Burnham, K. P. & Anderson, D. R. (1998). Practical use of the information-theoretic approach. In *Model selection and inference* (pp. 75–117). Springer.

[31] Byrne, J. E., Hughes, M. E., Rossell, S. L., Johnson, S. L., & Murray, G. (2017). Time of day differences in neural reward functioning in healthy young men. *Journal of Neuroscience*, (pp. 0918–17).

[32] Caballero, J. A., Humphries, M. D., & Gurney, K. N. (2018). A probabilistic, distributed, recursive mechanism for decision-making in the brain. *PLoS Comput. Biol.*, 14(4), e1006033.

[33] Carrillo-Reid, L., Hernandez-Lopez, S., Tapia, D., Galarraga, E., & Bargas, J. (2011). Dopaminergic Modulation of the Striatal Microcircuit: Receptor-Specific Configuration of Cell Assemblies. *Journal of Neuroscience*, 31(42), 14972–14983.

[34] Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, 143(4), 1476.

[35] Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *Elife*, 9, e51260.

[36] Collins, A. G. E. & Frank, M. J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.*, 121(3), 337–366.

[37] Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292), 1464–1468.

[38] Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4), 613–624.

[39] Dayan, P. & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, 18(2), 185–196.

[40] Dayan, P. & Yu, A. J. (2006). Phasic norepinephrine: a neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, 17(4), 335–350.

[41] de Cothi, W., Nyberg, N., Griesbauer, E.-M., Ghanamé, C., Zisch, F., Lefort, J., Fletcher, L., Newton, C., Renaudineau, S., Bendor, D., et al.

(2020). Predictive maps in rats and humans for spatial navigation. *bioRxiv*.

[42] Diedrichsen, J. & Shadmehr, R. (2005). Detecting and adjusting for artifacts in fMRI time series data. *NeuroImage*, 27(3), 624–634.

[43] Dunovan, K., Lynch, B., Molesworth, T., & Verstynen, T. (2015). Competing basal ganglia pathways determine the difference between stopping and deciding not to go. *Elife*, 4, e08723.

[44] Dunovan, K. & Verstynen, T. (2016). Believer-skeptic meets actor-critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. *Frontiers in neuroscience*, 10, 106.

[45] Dunovan, K. & Verstynen, T. (2019). Errors in action timing and inhibition facilitate learning by tuning distinct mechanisms in the underlying decision process. *Journal of Neuroscience*, 39(12), 2251–2264.

[46] Dunovan, K., Vich, C., Clapp, M., Verstynen, T., & Rubin, J. (2019). Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making. *PLoS computational biology*, 15(5), e1006998.

[47] Ekstrom, A. D. & Ranganath, C. (2018). Space, time, and episodic memory: The hippocampus is all over the cognitive map. *Hippocampus*, 28(9), 680–687.

[48] Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2018). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1), 111–116.

[49] Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature reviews neuroscience*, 9(4), 292–303.

[50] Feng, S. F., Wang, S., Zarnescu, S., & Wilson, R. C. (2020). The dynamics of explore-exploit decisions reveal a signal-to-noise mechanism for random exploration.

[51] Feng, S. F., Wang, S., Zarnescu, S., & Wilson, R. C. (2021). The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific reports*, 11(1), 1–15.

[52] Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature neuroscience*, 22(12), 2066–2077.

[53] Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., Bogacz, R., & Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proc. Natl. Acad. Sci. U. S. A.*, 107(36), 15916–15920.

[54] Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proc. Natl. Acad. Sci. U. S. A.*, 105(45), 17538–17542.

[55] Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual review of psychology*, 67.

[56] Foster, E. D. & Deardorff, A. (2017). Open science framework (osf). *Journal of the Medical Library Association: JMLA*, 105(2), 203.

[57] Friend, D. M. & Kravitz, A. V. (2014). Working together: basal ganglia pathways in action selection. *Trends in neurosciences*, 37(6), 301–3.

[58] Gammaitoni, L., Hänggi, P., Jung, P., & Marchesoni, F. (1998). Stochastic resonance. *Reviews of modern physics*, 70(1), 223.

[59] Gauthier, I. & Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision research*, 37(12), 1673–1682.

[60] Gershman, S. J. (2018). The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, 38(33), 7193–7200.

[61] Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, 6(3), 277.

[62] Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological review*, 117(1), 197.

[63] Gershman, S. J. & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, 20(2), 251–256.

[64] Gershman, S. J. & Tzovaras, B. G. (2018). Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia*, 120, 97–104.

[65] Gershman, S. J. & Uchida, N. (2019). Believing in dopamine. *Nature Reviews Neuroscience*, 20(11), 703–714.

[66] Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience*, 10(2), 252–269.

[67] Gold, J. I. & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, 30(30), 535–561.

[68] Gureckis, T. M. & Love, B. C. (2009). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, 53(3), 180–193.

[69] Gurney, K., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological cybernetics*, 84(6), 411–23.

[70] Gurney, K. N., Humphries, M. D., & Redgrave, P. (2015). A new framework for cortico-striatal plasticity: behavioural theory meets in vitro data at the reinforcement-action interface. *PLoS biology*, 13(1), e1002034.

[71] Herz, D. M., Tan, H., Brittain, J.-S., Fischer, P., Cheeran, B., Green, A. L., FitzGerald, J., Aziz, T. Z., Ashkan, K., Little, S., Foltynie, T., Limousin, P., Zrinzo, L., Bogacz, R., & Brown, P. (2017). Distinct mechanisms mediate speed-accuracy adjustments in cortico-subthalamic networks. *Elife*, 6.

[72] Herz, D. M., Zavala, B. A., Bogacz, R., & Brown, P. (2016a). Neural correlates of decision thresholds in the human subthalamic nucleus. *Current Biology*, 26(7), 916–920.

[73] Herz, D. M., Zavala, B. A., Bogacz, R., & Brown, P. (2016b). Neural correlates of decision thresholds in the human subthalamic nucleus. *Current Biology*, 26(7), 916–920.

[74] Heston, J., Friedman, A., Baqai, M., Bavafa, N., Aron, A. R., & Hnasko, T. S. (2020). Activation of subthalamic nucleus stop circuit disrupts cognitive performance. *eNeuro*.

[75] Hollerman, J. R. & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.*, 1(4), 304–309.

[76] Hurwicz, L. (1972). On informationally decentralized systems. *Decision and Organization*, (pp. 320).

[77] Jahfari, S., Ridderinkhof, K. R., Collins, A. G., Knapen, T., Waldorp, L. J., & Frank, M. J. (2019). Cross-task contributions of frontobasal ganglia circuitry in response inhibition and conflict-induced slowing. *Cerebral Cortex*, 29(5), 1969–1983.

[78] Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

[79] Jepma, M. & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory. *Journal of cognitive neuroscience*, 23(7), 1587–1596.

[80] Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, 89(1), 221–234.

[81] Kakade, S. & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, 15(4-6), 549–559.

[82] Kaplan, R., Schuck, N. W., & Doeller, C. F. (2017). The role of mental maps in decision-making. *Trends in Neurosciences*, 40(5), 256–259.

[83] Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31), 10687–10692.

[84] Keung, W., Hagen, T. A., & Wilson, R. C. (2019). Regulation of evidence accumulation by pupil-linked arousal processes. *Nature Human Behaviour*, 3(6), 636–645.

[85] Kitano, H., Tanibuchi, I., & Jinnai, K. (1998). The distribution of neurons in the substantia nigra pars reticulata with input from the motor, premotor and prefrontal areas of the cerebral cortex in monkeys. *Brain Research*, 784(1-2), 228–238.

[86] Klaus, A., Martins, G. J., Paixão, V. B., Zhou, P., Paninski, L., & Costa, R. M. (2017). The spatiotemporal organization of the striatum encodes action space. *Submitted*, 95(5), 1171–1180.e7.

[87] Kloosterman, N. A., de Gee, J. W., Werkle-Bergner, M., Lindenberger, U., Garrett, D. D., & Fahrenfort, J. J. (2019). Humans strategically shift decision bias by flexibly adjusting sensory evidence accumulation. *Elife*, 8, e37321.

[88] Kropotov, J. D. & Etlinger, S. C. (1999). Selection of actions in the basal ganglia–thalamocortical circuits: Review and model. *International Journal of Psychophysiology*, 31(3), 197–217.

[89] Kruschke, J. K. & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. *The Oxford handbook of computational and mathematical psychology*, (pp. 279–299).

[90] Ledyard, J. O. (1989). Incentive compatibility. In *Allocation, Information and Markets* (pp. 141–151). Springer.

[91] Lee, J., Wang, W., & Sabatini, B. L. (2020). Anatomically segregated basal ganglia pathways allow parallel behavioral modulation. *Nature Neuroscience*.

[92] Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. (2019). Human replay spontaneously reorganizes experience. *Cell*, 178(3), 640–652.

[93] Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in r. *Behavior research methods*, 49(4), 1494–1502.

[94] Machado, M. C., Bellemare, M. G., & Bowling, M. (2017a). A laplacian framework for option discovery in reinforcement learning. *arXiv preprint arXiv:1703.00956*.

[95] Machado, M. C., Rosenbaum, C., Guo, X., Liu, M., Tesauro, G., & Campbell, M. (2017b). Eigenoption discovery through the deep successor representation. *arXiv preprint arXiv:1710.11089*.

[96] Mark, S., Moran, R., Parr, T., Kennerley, S. W., & Behrens, T. E. (2020). Transferring structural knowledge across cognitive maps in humans and models. *Nature Communications*, 11(1), 1–12.

[97] McClure, S. M., Gilzenrat, M. S., & Cohen, J. D. (2005). An exploration-exploitation model based on norepinepherine and dopamine activity. *Advances in neural information processing systems*, 18, 867–874.

[98] McDonnell, M. D. & Ward, L. M. (2011). The benefits of noise in neural systems: bridging theory and experiment. *Nature Reviews Neuroscience*, 12(7), 415–425.

[99] Mendonça, A. G., Drugowitsch, J., Vicente, M. I., DeWitt, E. E., Pouget, A., & Mainen, Z. F. (2020). The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. *Nature Communications*, 11(1), 1–15.

[100] Mikhael, J. G. & Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. *PLoS Comput. Biol.*, 12(9), e1005062.

[101] Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Prog. Neurobiol.*, 50(4), 381–425.

[102] Momennejad, I. (2020). Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, 32, 155–166.

[103] Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2017a). Offline replay supports planning: fmri evidence from reward revaluation. *bioRxiv*, (pp. 196758).

[104] Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *Elife*, 7, e32548.

[105] Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017b). The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9), 680–692.

[106] Mulder, K. & Klugkist, I. (2017). Bayesian estimation and hypothesis tests for a circular generalized linear model. *Journal of mathematical psychology*, 80, 4–14.

[107] Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. (2011). Pupillometry and p3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology*, 48(11), 1532–1543.

[108] Murphy, P. R., Vandekerckhove, J., & Nieuwenhuis, S. (2014). Pupil-linked arousal determines variability in perceptual decision making. *PLoS computational biology*, 10(9), e1003854.

[109] Murray, G., Nicholas, C. L., Kleiman, J., Dwyer, R., Carrington, M. J., Allen, N. B., & Trinder, J. (2009). Nature's clocks and human mood: The circadian system modulates reward motivation. *Emotion*, 9(5), 705.

[110] Nambu, A. (2011). Somatotopic Organization of the Primate Basal Ganglia. *Frontiers in Neuroanatomy*, 5(April), 1–9.

[111] Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*, 15(7), 1040.

[112] Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 30(37), 12366–12378.

[113] Neumann, W. J., Schroll, H., De Almeida Marcelino, A. L., Horn, A., Ewert, S., Irmen, F., Krause, P., Schneider, G. H., Hamker, F., & Kühn, A. A. (2018). Functional segregation of basal ganglia pathways in Parkinson's disease. *Brain*, 141(9), 2655–2669.

[114] Nguyen, K. P., Josić, K., & Kilpatrick, Z. P. (2019). Optimizing sequential decisions in the drift–diffusion model. *Journal of mathematical psychology*, 88, 32–47.

[115] O'keefe, J. & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

[116] O'reilly, J. X. (2013). Making predictions in a changing world—inference, uncertainty, and learning. *Frontiers in neuroscience*, 7, 105.

[117] Payzan-LeNestour, E. & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput Biol*, 7(1), e1001048.

[118] Payzan-LeNestour, É. & Bossaerts, P. (2012). Do not bet on the unknown versus try to find out more: estimation uncertainty and "unexpected uncertainty" both modulate exploration. *Frontiers in neuroscience*, 6, 150.

[119] Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic bulletin & review*, 24(4), 1234–1251.

[120] Pfeiffer, B. E. (2020). The content of hippocampal "replay". *Hippocampus*, 30(1), 6–18.

[121] Pfeiffer, B. E. & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447), 74–79.

[122] Pisupati, S., Chartarifsky-Lynn, L., Khanal, A., & Churchland, A. K. (2021). Lapses in perceptual decisions reflect exploration. *Elife*, 10, e55490.

[123] Ponzi, A. & Wickens, J. (2010). Sequentially switching cell assemblies in random inhibitory networks of spiking neurons in the striatum. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(17), 5894–911.

[124] Prat-Carrabin, A., Wilson, R. C., Cohen, J. D., & Da Silveira, R. A. (2020). Human inference in changing environments with temporal structure. *BioRxiv*, (pp. 720516).

[125] Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: phasic and tonic changes are associated with altered vigilance. *Brain research bulletin*, 35(5-6), 607–616.

[126] Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, 85(2), 59.

[127] Ratcliff, R. & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural computation*, 24(5), 1186–1229.

[128] Ratcliff, R. & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4), 873–922.

[129] Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature communications*, 7(1), 1–7.

[130] Rubin, J. E., Vich, C., Clapp, M., Noneman, K., & Verstynen, T. (2020). The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution. *European Journal of Neuroscience*.

[131] Rubin, J. E., Vich, C., Clapp, M., Noneman, K., & Verstynen, T. (2021). The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution. *European Journal of Neuroscience*, 53(7), 2234–2253.

[132] Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS computational biology*, 13(9), e1005768.

[133] Sadeghiyeh, H., Wang, S., Alberhasky, M. R., Kyllo, H. M., Shenhav, A., & Wilson, R. C. (2020). Temporal discounting correlates with directed exploration but not with random exploration. *Scientific reports*, 10(1), 1–10.

[134] Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6), 110–114.

[135] Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci.*, 12(12), 4595–4610.

[136] Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, 116(28), 13903–13908.

[137] Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive psychology*, 119, 101261.

[138] Schulz, E. & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, 55, 7–14.

[139] Shan, Q., Ge, M., Christie, M. J., & Balleine, B. W. (2014). The acquisition of goal-directed actions generates opposing plasticity in direct and indirect pathways in dorsomedial striatum. *Journal of Neuroscience*, 34(28), 9196–9201.

[140] Sirois, S. & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(6), 679–692.

[141] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4), 583–639.

[142] Spiers, A. & Calne, D. (1969). Action of dopamine on the human iris. *Br Med J*, 4(5679), 333–335.

[143] Stein, R. B., Gossen, E. R., & Jones, K. E. (2005). Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience*, 6(5), 389–397.

[144] Stoianov, I., Maisto, D., & Pezzulo, G. (2020). The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning. *bioRxiv*.

[145] Stolle, M. & Precup, D. (2002). Learning options in reinforcement learning. In *International Symposium on abstraction, reformulation, and approximation* (pp. 212–223).: Springer.

[146] Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, 25(3), 251–260.

[147] Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4), 160–163.

[148] Sutton, R. S. & Barto, A. G. (1998). *Introduction to Reinforcement Learning*. Camgridge: MIT Press.

[149] Sutton, R. S. & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[150] Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature neuroscience*, 22(9), 1503–1511.

[151] Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, 55(4), 189.

[152] Uddin, L. Q. (2020). Bring the noise: reconceptualizing spontaneous neural activity. *Trends in cognitive sciences*, 24(9), 734–746.

[153] Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature communications*, 8(1), 1–11.

[154] Urai, A. E., De Gee, J. W., & Donner, T. H. (2018). Choice history biases subsequent evidence accumulation. *BioRxiv*, (pp. 251595).

[155] Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity reveals a novel dissociation between action and confidence. *Neuron*, 96(2), 348–354.

[156] van Kempen, J., Loughnane, G. M., Newman, D. P., Kelly, S. P., Thiele, A., O'Connell, R. G., & Bellgrove, M. A. (2019). Behavioural and neural signatures of perceptual decision-making are modulated by pupil-linked arousal. *Elife*, 8, e42541.

[157] Vich, C., Clapp, M., Verstynen, T., & Rubin, J. E. (2022). Identifying control ensembles for information processing within the cortico-basal ganglia-thalamic circuit. *bioRxiv*.

[158] Vich, C., Dunovan, K., Verstynen, T., & Rubin, J. (2019). Corticostriatal synaptic weight evolution in a two-alternative forced choice task. *bioRxiv*.

[159] Vich, C., Dunovan, K., Verstynen, T., & Rubin, J. (2020). Corticostriatal synaptic weight evolution in a two-alternative forced choice task: a computational study. *Communications in Nonlinear Science and Numerical Simulation*, 82, 105048.

[160] Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review*, 14(5), 779–804.

[161] Waltz, J. A., Wilson, R. C., Albrecht, M. A., Frank, M. J., & Gold, J. M. (2020). Differential effects of psychotic illness on directed and random exploration. *Computational Psychiatry*, 4, 18–39.

[162] Warren, C. M., Wilson, R. C., Van Der Wee, N. J., Giltay, E. J., Van Noorden, M. S., Cohen, J. D., & Nieuwenhuis, S. (2017). The effect of atomoxetine on random and directed exploration in humans. *PloS one*, 12(4), e0176034.

[163] Wei, W., Rubin, J. E., & Wang, X.-J. (2015). Role of the indirect pathway of the basal ganglia in perceptual decision making. *Journal of Neuroscience*, 35(9), 4052–4064.

[164] Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*.

[165] Wiecki, T. V. & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological review*, 120(2), 329–55.

[166] Wiecki, T. V., Sofer, I., & Frank, M. J. (2013a). Hddm: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, 7, 14.

[167] Wiecki, T. V., Sofer, I., & Frank, M. J. (2013b). Hddm: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, 7, 14.

[168] Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2020). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, 38, 49–56.

[169] Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, 38, 49–56.

[170] Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6), 2074.

[171] Wilson, R. C., Nassar, M. R., & Gold, J. I. (2010). Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, 22(9), 2452–2476.

[172] Wilson, R. C. & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in human neuroscience*, 5, 189.

[173] Wu, X. & Foster, D. J. (2014). Hippocampal replay captures the unique topological structure of a novel environment. *Journal of Neuroscience*, 34(19), 6459–6469.

[174] Yartsev, M. M., Hanks, T. D., Yoon, A. M., & Brody, C. D. (2018). Causal contribution and dynamical encoding in the striatum during evidence accumulation. *Elife*, 7, e34929.

[175] Yerkes, R. M., Dodson, J. D., et al. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Punishment: Issues and experiments*, (pp. 27–41).

[176] Yttri, E. A. & Dudman, J. T. (2016). Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature*, 533(7603), 1–16.

[177] Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife*, 6, e27430.