Dynamic decision policy reconfiguration under outcome uncertainty

Krista Bond[1,2,3*], **Kyle Dunovan**[4], **Alexis Porter**[5], **Jonathan Rubin**[2,6], **Timothy Verstynen**[1,2,3,7*]

***For correspondence:**
kbond@andrew.cmu.edu (KB);
timothyv@andrew.cmu.edu (TV)

[1]Department of Psychology, Carnegie Mellon University, United States; [2]Center for the Neural Basis of Cognition, United States; [3]Carnegie Mellon Neuroscience Institute, United States; [4]Ntrepid Corporation, United States; [5]Department of Psychology, Northwestern University, United States; [6]Department of Mathematics, University of Pittsburgh, United States; [7]Department of Biomedical Engineering, Carnegie Mellon University, United States

Abstract Humans and other mammals flexibly select actions under noisy and unstable conditions. To shed light on the mechanism driving this flexibility, we evaluated how the underlying decision policy evolves when humans change their minds about the most rewarding action. Participants performed a dynamic variant of the two-armed bandit task that manipulated the certainty in relative reward probabilities (conflict) and the reliability of action-outcome contingencies (volatility). We found that conflict and volatility contributed to shifts in exploratory states by changing both the rate of evidence accumulation (drift rate) and the amount of evidence needed to make a decision (boundary height). Following a switch in the optimal choice, the drift rate and the boundary height reduce, allowing variability in the accumulation process to predominate action selection, leading to a fast exploratory state. These changes facilitate the discovery of the new optimal choice, with a quick recovery of the boundary height to baseline. In parallel, the drift rate gradually returns to its asymptotic value as the belief in the value of the optimal choice stabilizes. Together, these decision dynamics suggest that, in the context of volatile two-choice decisions, humans adopt a combined information-threshold and drift rate mechanism in response to environmental changes. Unlike previous observations, we found no evidence that fluctuations in norepinephrine, as measured by pupillometry, associated with this adaptive shift toward an exploratory policy. We conclude that the multifaceted processes underlying a decision can rapidly reconfigure to adapt action selection policy under multiple forms of environmental uncertainty.

Introduction

Humans are faced with uncertain choices on a daily basis. 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 dilemma: 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) (***O'reilly, 2013***).

Navigating the exploration-exploitation dilemma depends on the environmental context. To illustrate this, Figure 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 (***Sutton and Barto*** (***1998***); 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) was independently manipulated. For each level of conflict and volatility, a set of tabular Q-learning (***Sutton and Barto, 1998***) 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 reward 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. 1A).

As the value of the optimal choice decreases relative to the value of a suboptimal choice (conflict increases), the learner tends to exploit what she already knows. But as action values grow unstable (volatility increases) when the clarity of the optimal choice is constant (constant conflict), the learner is biased toward exploration (***Bland et al., 2012***). As these two forms of uncertainty change together, the gradient of action selection strategy also changes. These simulations highlight the multivariate contributions to optimal action selection policy.

Less certain environments prompt a more frequent change of mind regarding the best option, and thus they are conducive to more exploratory decision policies. How do underlying decision processes change to reflect this shift toward exploration after detecting a possible change in the optimal choice? One popular set of algorithms for describing the dynamics of decision making are accumulation-to-bound processes like the drift-diffusion model (DDM; ***Ratcliff*** (***1978***)). 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. 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). Decreasing both drift rate and boundary height leads to more random decisions (Fig. 1C), however, the speed of these decisions depends on the ratio of the two parameters (Fig. 1D). Thus exploratory policies can result in either fast or slow decisions, depending on the relative configuration of drift rate and boundary height.

Previous modeling work has shown that these properties of a DDM process can be modulated by feedback signals and choice history (***Pedersen et al., 2017***; ***Ratcliff and Frank, 2012***; ***Dunovan and Verstynen, 2019***; ***Dunovan et al., 2019***; ***Mendonça et al., 2020***; ***Urai et al., 2018***) with different mechanisms for adapting the drift rate and the boundary height. When 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 (***Dunovan et al., 2019***; ***Mikhael and Bogacz, 2016***; ***Bariselli et al., 2018***; ***Rubin et al., 2020***). In contrast to value estimation, selection errors have been linked to changes in the boundary height (***Forstmann et al., 2008***, ***2010***; ***Bogacz et al., 2010***; ***Herz et al., 2016a***, ***2017***; ***Dunovan et al., 2019***; ***Dunovan and Verstynen, 2019***) and internal estimates of environmental change (***Nassar et al., 2010***; ***Wilson and Niv, 2012***; ***Nassar et al., 2012***; ***Behrens et al., 2007***). Given the adaptive sensitivity of the drift rate and the boundary height to value estimation and selection, respectively, these decision parameters may define unique states on a manifold of fast or slow and exploratory or exploitative decision policies. These policies, in turn, adaptively reconfigure based on current environmental feedback signals by modulating value estimation and the rate of selection errors (Fig. 1E). Agents can move along the decision manifold from exploitative states (bright colors, Fig. 1E) to different types of exploratory states (darker colors, Fig. 1E) as they commit a greater number of selection errors prompted by change in action-outcome contingencies. As the system relearns properties of the environment, the decision policy migrates along the manifold to return to an exploitative state until a change occurs again.

One plausible mechanism for this migration along the manifold of selection policies is the locus coereleus norepinephrine (LC-NE) system, which has been linked to adaptive behavioral variability in response to uncertainty (***Urai et al., 2017***; ***Dayan and Yu, 2006***; ***Bouret and Sara, 2005***). The LC-

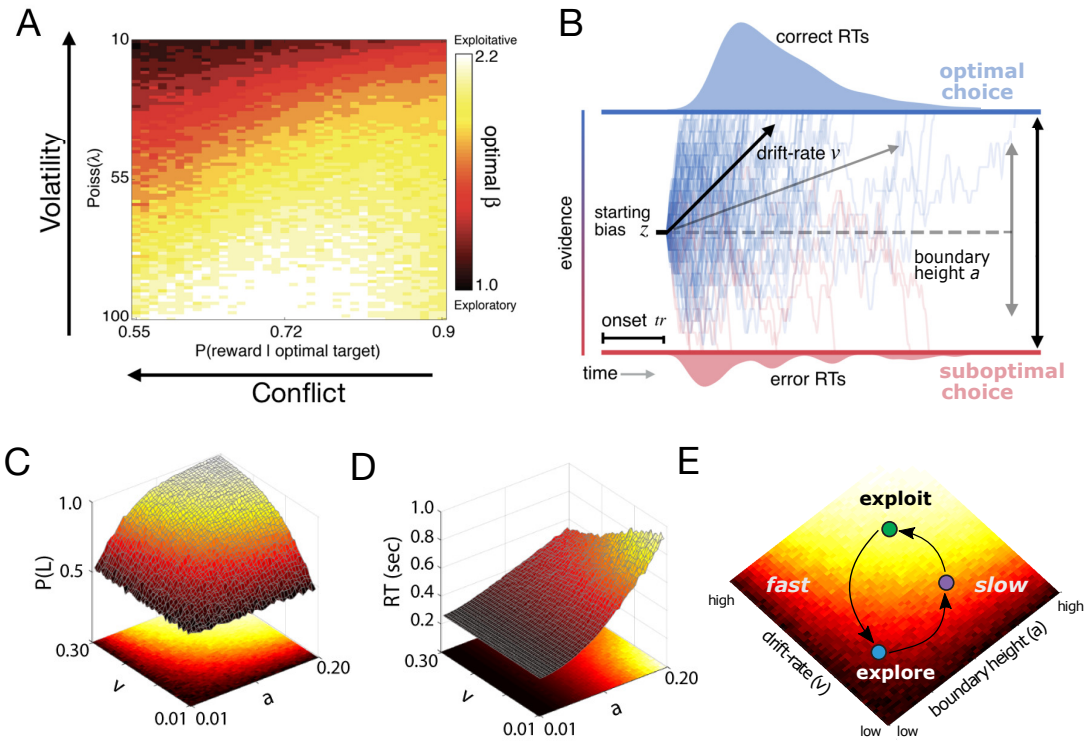


Figure 1. A) The degree of conflict and volatility shifts the optimal balance between exploration and exploitation. B) The drift diffusion model. C) Accuracy (probability that left choice selected is selected; P(L)) as a function of coordinated changes in the rate of evidence accumulation (v) and the 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.

NE system has two distinct modes (***Aston-Jones and Bloom, 1981***) that map onto distinct decision states (***Aston-Jones and Cohen, 2005***). 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 (***Aston-Jones and Bloom, 1981***). 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 (***Aston-Jones and Cohen, 2005***). Thus the LC-NE system, which can be indirectly measured by fluctuations in pupil diameter (***Aston-Jones and Cohen, 2005***; ***Jepma and Nieuwenhuis, 2011***), may be a central mechanism for modulating selection policies.

Here we investigated the dynamics of decision policy changes in response to different types of outcome uncertainty in a series of experiments. To shift the greediness of the selection policy, we varied the ambiguity in optimal choice (*conflict*) and the instability of action outcomes (*volatility*) as human participants performed a dynamic variant (***Daw et al., 2006***) of the two-armed bandit task (***Sutton and Barto, 2018***). We predicted that, in response to suspected changes in action-outcome relationships, humans would exhibit a stereotyped adjustment in the drift rate and boundary height, that changes in response to different types of uncertainty. In addition, using pupillary data, we explored whether the LC-NE system covaried with the decision dynamics that we previously observed. We predicted that fluctuations in norepinephrine would shift the boundary height in response to a change in action outcomes to facilitate exploration.

Results

Across two experiments we used a two-armed bandit task with equivalent sensory uncertainty between targets to independently manipulate the reward conflict and the volatility of action outcomes in order to measure how underlying decision processes responded to environmental perturbations (see 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, we sought to establish the effect of conflict and volatility on latent decision processes. Twenty-four participants completed four sessions (high and low conflict; high and low volatility) composed of 600 trials each. During each session, they were asked to select one of two coin boxes (Exp. 1: Fig. 2A). The levels of conflict and volatility for all four conditions in Experiment 1 are shown as gray dots in Fig. 2C. Experiment 2 was an out-of-set validation of the effects that we observed in Experiment 1. 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. 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 nine conditions in Experiment 2 are shown as black dots in Fig. 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 results for both of these experiments are presented together below.

The influence of ambiguity and instability on speed and accuracy

We first considered the effect of our conditional manipulations on the speed and accuracy of choices for 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. 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 reaction times (Fig. 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. 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 ~ 13 ms on average; Fig. 3B).

Experiment 2 served as a high powered test of whether the effects we observed in Experiment 1 were replicable. In this experiment, we found similar effects of conflict and volatility on accuracy as we observed in Experiment 1. Accuracy increased as conflict decreased (i.e. as the probability of reward increased; β=1.312, 95% CI=1.259,1.368, z=12.752, p<2e-16). As the environment grew less volatile, accuracy increased (β=1.131, 95% CI=1.085,1.179, z=5.798, p=6.72e-09). Unlike in Experiment 1, reaction times in Experiment 2 reliably increased with the probability of reward (β=0.022, 95% CI=0.021,0.024, t=23.430, p<2e-16). Reaction time also *decreased* as volatility increased (i.e., as epoch length decreased; Fig. 3D; β=0.006, 95% CI=0.004,0.008, t=5.966, p=2.49e-09).

In light of the data from both of these experiments, we conclude that increasing the ambiguity of the optimal choice (conflict) and increasing the instability of action outcomes (volatility) decreases the probability of selecting the optimal choice. We observed inconsistent effects of conflict and

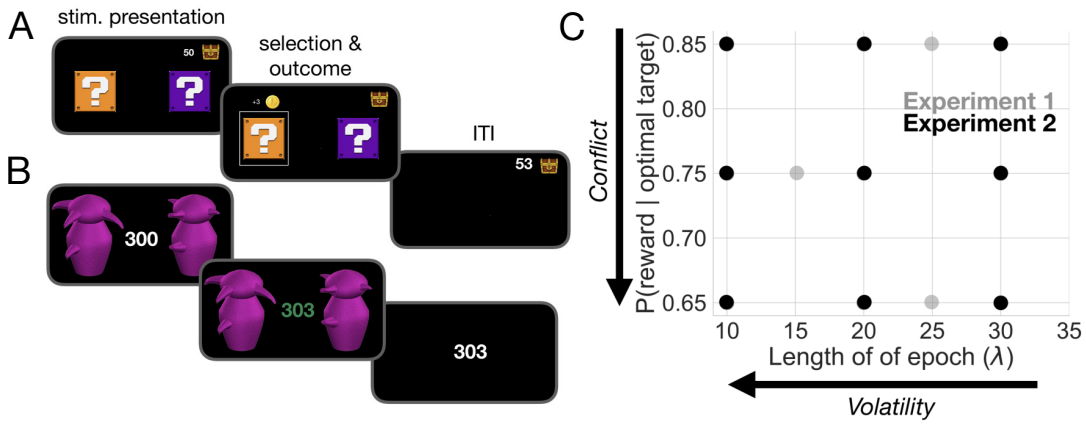


Figure 2. Task and uncertainty manipulation. A) In Experiment 1, participants were asked to choose between one of two "mystery boxes". The point value associated with a selection was displayed above the chosen mystery box. The sum of points earned across trials was shown to the left of a treasure box 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. 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.

volatility on reaction times across experiments. While neither conflict nor volatility had a strong effect on reaction times in Experiment 1, Experiment 2 suggests a slight but statistically reliable *increase* in reaction time as choices become less ambiguous and a *decrease* in reaction time as the environment grows less stable. This suggests that as the time horizon for finding the optimal selection shrinks, the speed of decisions accelerates. Independently, as the identity of the optimal choice becomes clearer, decisions take more time.

Tracking estimates of action value and environmental volatility

We calculated trial-by-trial estimates of two ideal observer parameters of environmental states that varied as a function of our manipulations of conflict and volatility (see Cognitive model for calculation details; ***Nassar et al.*** (***2010***); ***Vaghi et al.*** (***2017***)). 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 complement of belief as doubt, such that when belief decreases doubt increases. We predicted that ΔB would capture the degree of conflict that participants estimated, with the difference in belief between these choices decreasing as conflict increases and increasing as conflict decreases. 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). We predicted that change point probability would capture the degree of volatility that participants estimated, increasing with the frequency of a switch in the identity of the optimal choice. 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. 1).

Consistent with our assumption that increasing conflict for the optimal choice should increase the doubt in the value of that choice, in Experiment 1 we observed a sharp decrease in ΔB after a switch in action outcomes and a gradual return to asymptotic values (Fig. 4A) with a decreased difference in reward probability resulting in increased doubt (4B; $\beta = 0.216$,95% CI:0.206, 0.224, t=46.24, p<2e-16). We found that both high and low volatility conditions resulted in a similar time course of belief updating relative to a switch in action outcomes (Fig. 4A). As expected, stable 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

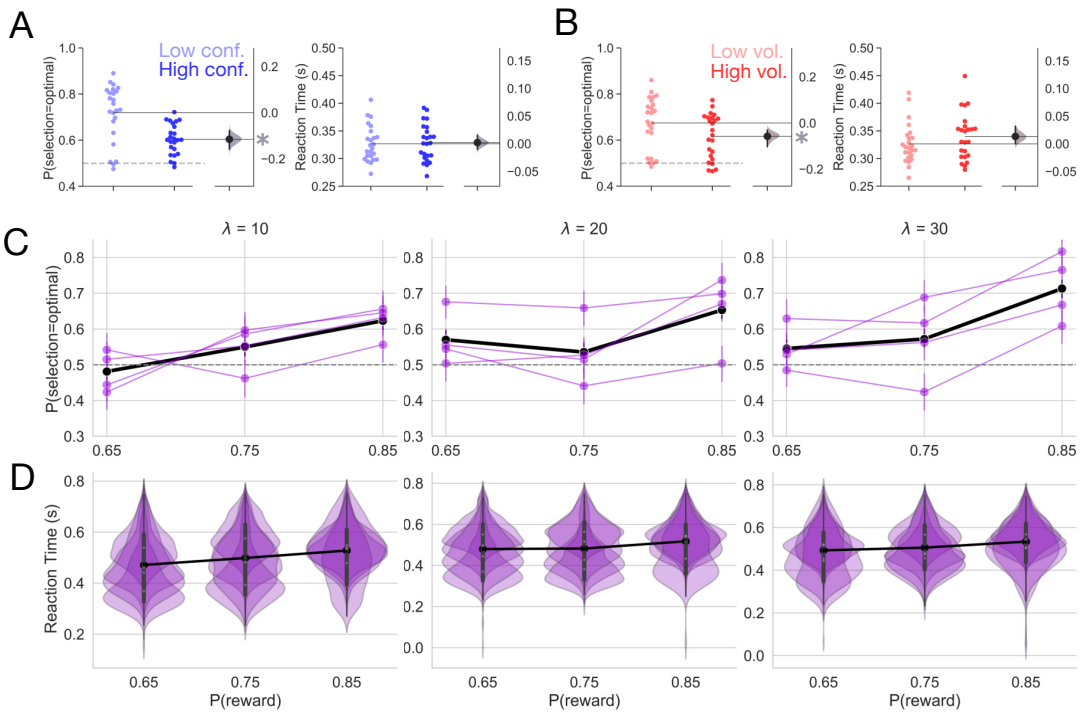


Figure 3. Behavior. A) Mean accuracy and reaction time for the manipulation of conflict in Experiment 1. B) Mean accuracy and reaction time for the manipulation of volatility in Experiment 1. Each point represents the average for a single subject. The distribution to the right represents the bootstrapped 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.

low conflict conditions allowed (see Fig. 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. 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. 4C). Correspondingly, over all trials, Ω was higher under more volatile conditions (Fig. 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. 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 (Fig. 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. 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 indeed

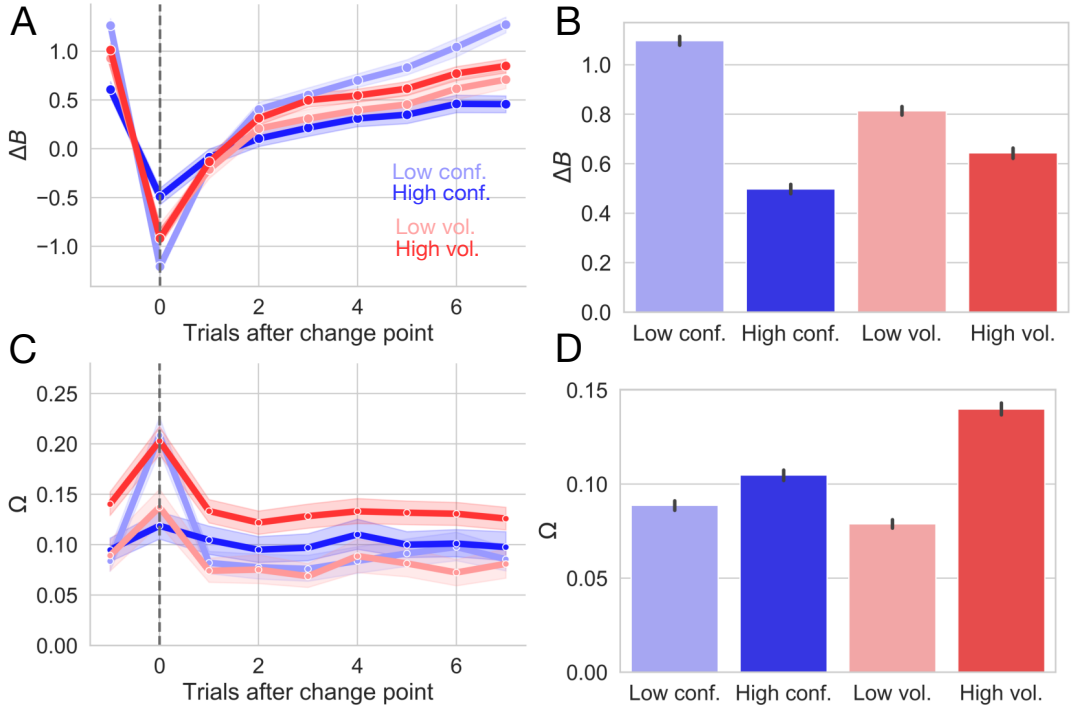


Figure 4. Changes in ideal observer estimates as a function of condition for Experiment 1. A) Changes in the belief in the value of the optimal target (ΔB) as a function of conflict and volatility over time. B) Mean belief in the value of the optimal choice across conditions. C) Changes in change point probability (Ω) as a function of conflict and volatility over time. D) Mean change point probability across conditions. Error bars represent 95% CIs.

more frequent (see increased Ω estimates for high volatility conditions over high conflict conditions in Figure 4D). These results suggest that our formulation of these ideal observer estimates adequately captures our manipulation of volatility and conflict at a continuous level.

We now consider the link between these internal estimates of uncertainty, ΔB (or doubt, the inverse of ΔB) and Ω and the behavioral effects we observed in Experiments 1 and 2. In Experiment 1, we found no substantial association between doubt and reaction time ($\beta = -0.003$, 95% CI: -0.005, 0.000, t=-2.45, p=0.015). Similarly, a one standard deviation increase in Ω did not reliably relate to reaction time to a meaningful degree ($\beta = 0.002$, 95% CI: 0.000, 0.005, t=2.15, p=0.032). Further, Ω and ΔB did not interact to affect reaction time ($\beta = -0.001$, 95% CI: -0.003, 0.001, t=-1.24, p=0.215). These results are consistent with the null effects of our conditional manipulations on reaction time (see The influence of ambiguity and instability on speed and accuracy). However, Ω and ΔB did interact to affect accuracy ($\beta = 0.916$, 95% CI: 0.905, 0.927, z=-14.51, p<2e-16), with the effect of belief on accuracy decreasing as the probability of a change point increased, consistent with the conditional effects of conflict and volatility decreasing the accuracy of selection.

We found similar results in Experiment 2. Accuracy increased as doubt decreased, or as the belief in the value of the optimal choice increased (Fig. 3C; β=1.156, 95% CI=1.126,1.186, z=10.884, p<2e-16). As the estimate of Ω increased, accuracy decreased (Fig. 3C; β=0.899, 95% CI=0.869,0.930, z=-6.220, p=4.97e-10). As observed in our analysis of the behavioral data in terms of our conditional manipulations, reaction times in Experiment 2 reliably increased to a small degree with ΔB (Fig. 3D; β=0.004, mean 95% CI=0.002,0.005, t=6.028, p=1.7e-09), unlike in Experiment 1. Consistent with our previously observed effect of volatility on reaction time in Experiment 2, reaction time *decreased* as Ω increased (β=-0.004, 95% CI=-0.005,-0.002, t=-4.716, p=2.43e-06), although this decrease in reaction time was only on the order of 4 ms for a one standard deviation increase in Ω, on average.

Across experiments, we find that the relationship between ideal observer estimates and behavior mirrors the relationship between our conditional manipulations and behavior. Specifically, selection accuracy increased with the belief in the value of the optimal choice in both experiments, with an increase in the estimate of change point probability decreasing with the probability of selecting the optimal choice. As before, we observed inconsistent effects of the belief in the value of the optimal choice and change point probability on reaction times across experiments. While neither of these internal estimates of uncertainty had a reliable effect on reaction times in Experiment 1, Experiment 2 suggests a slight but statistically reliable *increase* in reaction time with the belief in the value of the optimal choice and a *decrease* in reaction time as the probability of a change increased.

Different forms of uncertainty impact distinct decision processes

Having evaluated how optimal choice ambiguity and unstable action outcomes affected gross behavior, we next moved on to how the underlying decision policy responded to these sources of uncertainty. To this end, we fit the behavioral data from Experiment 1 to versions of the adaptive DDM (Table 1, left panel) in which the ideal observer parameters explained trial-wise variability in either the rate of information accumulation (v), the amount of information needed to select an action (a), or both.

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 a high powered, within-subject test of whether the mapping between decision processes and ideal observer estimates that we observed in Experiment 1 would replicate. To select the model that best accounted for the data, we compared the deviance information criterion (DIC) scores (***Spiegelhalter et al., 2002***) for various models where the ideal observer parameters for environmental state (ΔB, Ω) covary with different parameters in the DDM process (Table 1). 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 (***Spiegelhalter et al., 2002***; ***Burnham and Anderson, 1998***).

Using the data from Experiment 1, we found that the model mapping ΔB to v and Ω to a provided the best model fit to the data given the models considered (minimum $\Delta_{DIC} = 13.1$) and fit the observed data substantially better than the null, intercept-only model ($\Delta_{DIC} = -2698$, see left panel of Table 1), which predicts the decision parameters (a and v) as a function of the mean values of those parameters. This suggests 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. An internal estimate of change point probability also updates the amount of information required to make a decision. To test whether the effects we observed in Experiment 1 were replicated in an independent data set, we performed a similar model comparison procedure in 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. For the sake of simplicity, we took the best fitting model from Experiment 1 (Model I) and compared this to the alternative model where the mapping between the ideal observer parameters and the DDM parameters is switched (Model II). This holds the complexity of our hypothesized and alternative model constant.

In three out of four participants in Experiment 2, we replicated the finding that the best model for explaining trialwise variability in DDM parameters was Model I (Table 1, right panel), with the DIC value for this model lower than the inverted model by an average of 31.8 points. Considering the model selection results from both experiments, we find robust evidence for an estimate of

Experiment 1

	ΔB	Ω	DIC	ΔDIC$_{null}$	ΔDIC$_{best}$
*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	DIC	ΔDIC$_{null}$	ΔDIC$_{best}$
*I	v	a	5496.6	-113.8	0.0
II	a	v	5596.0	-14.4	99.4
*I	v	a	5392.7	-42.3	0.0
II	a	v	5412.3	-22.8	19.6
*I	v	a	3157.5	-9.4	0.0
II	a	v	3162.7	-4.2	5.2
I	v	a	5456.3	-1.3	0.0
II	a	v	5459.2	1.6	2.9

Table 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 hypothesized mapping and the inverse mapping, with each set of models representing the fit for each of four subjects. 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.

belief in the value of the optimal choice changing the rate of information accumulation (v) and an estimate of change point probability changing the information threshold for making a decision (a).

To determine *whether* these internal estimates of uncertainty, ΔB and Ω, reliably affected the underlying decision processes, v and a, we examined the posterior distributions of the regression coefficients for the model supported across experiments (see Hierarchical drift diffusion modeling). We considered the relationship to be reliable if the estimated coefficient maintained the same sign over at least 95% of the mass of the posterior distribution. For Experiment 2, we performed the same analyses of the posterior distributions of regression coefficients and computed the mean and standard deviation of both the reliability of the estimates and the magnitude of the effect over subjects. For Experiment 1, both the relationship between a and Ω and the relationship between v and ΔB were statistically reliable (observed $p(\beta_{a\sim\Omega} > 0) = 0.002$; observed $p(\beta_{v\sim\Delta B} < 0) < 0.001$; see Supp. Fig. 2 for raw posterior distributions for Experiment 1).

Next, we asked *how* these internal estimates of uncertainty updated the decision processes. We found that the rate of evidence accumulation, v, increased with the belief in the value of the optimal choice, ΔB ($\beta_{v\sim\Delta B} = 0.404$, empirical 95% CI: 0.390, 0.418). Interestingly, we also observed that the amount of evidence needed to make a decision *decreased* as the probability of a change point increased ($\beta_{a\sim\Omega} = -0.035$, 95% CI: -0.057, -0.014).

Experiment 2 replicated the positive relationship between ΔB and v (mean observed $p(\beta_{v\sim\Delta B} < 0) = 0.003 \pm 0.005$), with the drift rate increasing as the belief in the value of the optimal choice increased (mean $\beta_{v\sim\Delta B} = 0.112 \pm 0.061$), though to a lesser degree than we observed originally (see Supp. Fig. 3 for raw posterior distributions for Experiment 2). The effect of Ω on a was less reliable than we observed in Experiment 1 (mean observed $p(\beta_{a\sim\Omega} < 0) = 0.727 \pm 0.108$) though the magnitude of the effect was on par with our previous observations (mean $\beta_{a\sim\Omega} = -0.037 \pm 0.017$), with the boundary height decreasing as the probability of a change point increased.

Across experiments, we find a consistent relationship between internal estimates of environmental uncertainty and the dynamics of the underlying decision processes. Altogether, these results suggest a combined information-threshold and drift rate mechanism for adaptation to change (Fig. 5A). When a change point is detected, the threshold for committing to a choice (a) decreases, potentiating a quick response. At the same time, the rate of evidence accumulation (v) approaches zero, allowing the variability in the decision process to dominate, encouraging a random selection. As the learner accrues information about the new optimal choice, the rate of information accumulation slowly recovers to asymptotic levels, with the decision process assuming a more directed path toward the choice that has accrued evidence for reward. Together, these independent, but cooperative, changes in the underlying decision processes point to a mechanism

for quickly gathering information in an unbiased manner when the learner needs to update her valuation. We now explore these dynamics in more detail.

Environmental instability prompts a stereotyped decision trajectory

So far we have established that both the drift rate and the boundary height are independently manipulated by two different estimates of environmental uncertainty with different temporal dynamics. This suggests that a change in action-outcome contingencies prompts a unique trajectory through the space of possible decision policies (Fig. 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. 5B). Following a shift in response contingencies, the navigation of this decision manifold follows a stereotyped pattern. The boundary height (a) and rate of evidence accumulation (v) initially decrease rapidly (see the trial prior to the change point in Fig. 5B) and then gradually recover and stabilize over time. This decision trajectory is consistent across subjects (Fig. 5B, right panel; see Supp. Fig. 4 for decision trajectories from Experiment 1).

To capture the consistency of this change-point driven response in the relationship between a and v over time, we also represented the trialwise shift in these two decision variables as vectors. The trial-by-trial estimates of drift rate and boundary height were calculated from the best-fit model and z-scored (see Different forms of uncertainty impact distinct decision processes for model selection). Then the difference between each sequential set of (a, v) coordinates 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. 5B, see Decision vector representation for methodological details). We find that the distribution of θ prior to a change point averages to $\sim 135\ \mathrm{deg}$, sharply changes in response to the observation of a change point ($\sim 270\ \mathrm{deg}$) and steadily returns to values prior to the onset of a change (Fig. 5C; see Supp. Fig. 5 for vector distributions). One trial after the change point, v sharply decreases along with a, after which a quickly recovers and v steadily progresses toward its baseline value. Altogether, 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 manifold of decision policies, we next asked whether conditions of increased volatility and increased conflict might modify its path. We conducted a comparison of null models and models specifying an influence of conflict and volatility using circular regression (***Mulder and Klugkist, 2017***). We began our model comparison with two null hypotheses. The first null model estimated decision dynamics (the relationship between a and v; θ) as a function of the intercept, or the average of the decision dynamics θ:

$$\theta = \beta_0$$

We call this the absolute null model.

Our second null model estimated decision dynamics (θ) solely as a function of time, with the time scale of consideration determined by the results of a stability analysis from Experiment 1 (see Model proposals and evaluation; Supp. Fig. 6):

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

We call this the time-null model.

Our model comparison logic was as follows. We first evaluated whether the posterior probability of the time-null model was greater than for the absolute null model. This would suggest that time was a better predictor of decision dynamics than the average response. If the 95% CI of the posterior probability for the time-null model exceeded the 95% CI of the absolute null model, we quantified the evidence for alternative models relative to the time-null model, with the sole effect of time as the new null hypothesis.

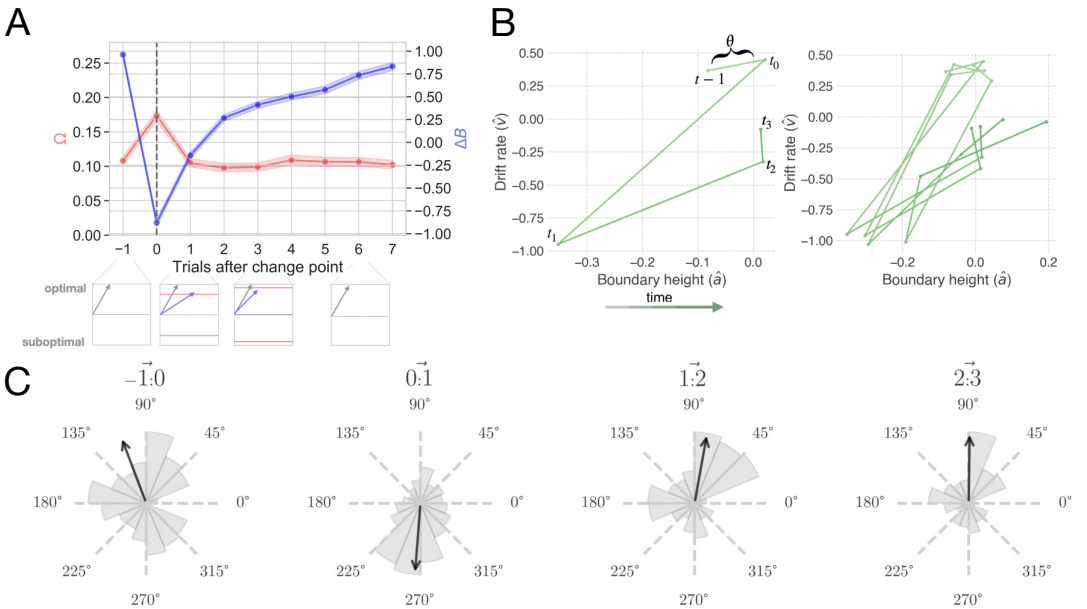


Figure 5. The decision manifold. A) Changes in ideal observer estimates of uncertainty over time and their effect on the boundary height and the drift rate. Directly after a change point, the boundary height *decreases* and the drift rate slows. Over time, the boundary height returns to its baseline value and the drift rate increases. B) Representing decision space in vector form. An angle was calculated between sequential values of (a,v) coordinates, beginning with the trial prior to the change point. The right-hand side shows the decision trajectories for all participants. Note that these trajectories are z-scored. C) Distributions depicting the angle between drift rate and boundary height. The area of the shaded region is proportional to the density and the arrow represents the circular mean.

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 ratio 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 (***Jeffreys, 1998***). Evidence for the null hypothesis is denoted BF_{01} and evidence for the alternative hypothesis is denoted BF_{10}. Because Experiment 2 took a replication-based approach, a separate model was fit for each participant for all proposed models.

To determine whether perturbations of volatility and conflict affected decision dynamics, we modeled decision dynamics (θ) 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. 6 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 (θ) by examining the posterior probability for each hypothesized model given the set of alternative hypotheses (Model proposals and evaluation; Fig. 6A). We found that the model describing the relationship between θ and time relative to a change point was more probable than the absolute null model (Fig. 6A).

We also present the evidence for the absolute null model against each alternative model as a Bayes Factor (BF_{01}) (Fig. 6B). The 95% confidence interval for the $BF01$ comparing the ratio of evidence for the absolute null model and the 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 absolute null model against the model specifying time-dependent effects of conflict included 1, suggesting no substantive difference between them. Given the equivocal evidence for the two models specifying time-dependent effects of conflict and volatility relative to the absolute null model, we excluded them from further comparison with the time-null model.

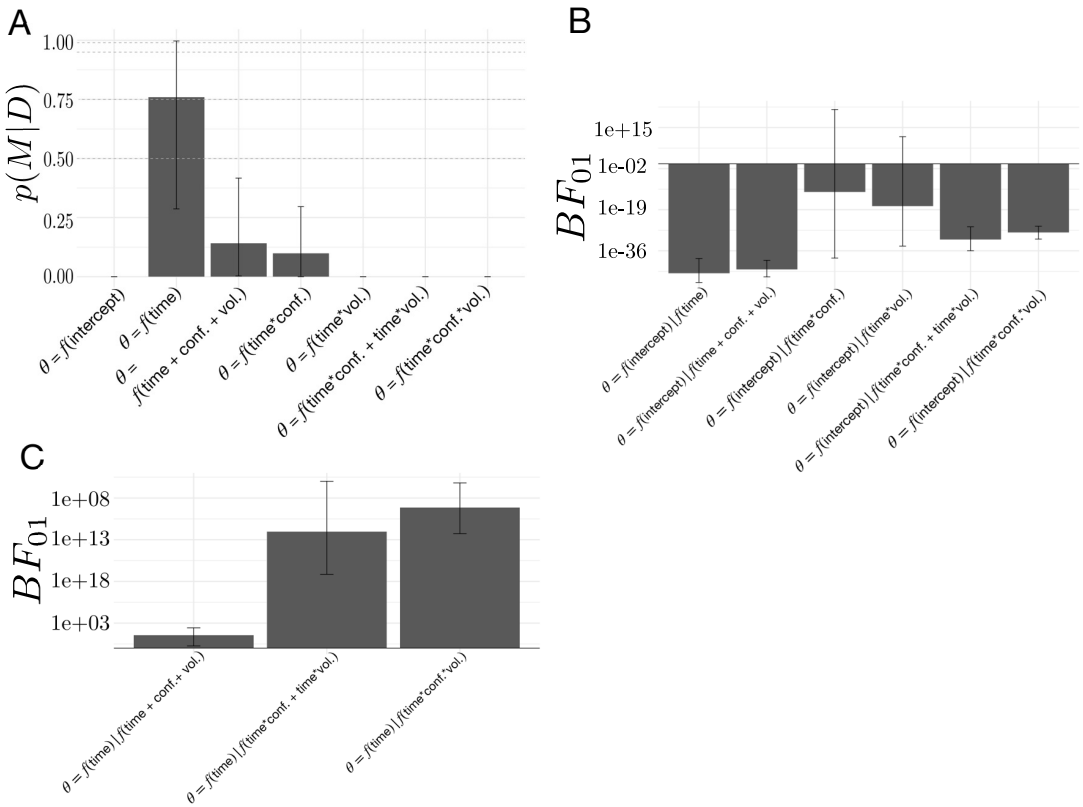


Figure 6. Model comparisons for the effect of volatility and conflict on the relationship between drift rate and boundary height (θ). A) The posterior probability for models testing for an effect of volatility and conflict on θ. B) The Bayes Factor for the intercept null model relative to the alternative models specifying either an effect of time alone or a conditional effect on θ. C) The Bayes Factor for the time null model relative to the surviving alternative models specifying a conditional effect on 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 (***Wagenmakers, 2007***).

The remainder of the models had substantially negative BF_{01}s (Fig. 6B), suggesting that they better fit the data than the absolute null model. To evaluate our second null hypothesis that time alone best accounted for the data, we computed the BF_{01} for the time-null model against the surviving models from the absolute null analysis. We find that, for all of the remaining models, the BF_{01} is substantially positive (Fig. 6C), indicating that the time-null model best accounted for the data (posterior probability of time-null 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 time relative to a change point. We find no evidence to suggest that changing the degree of volatility or changing the degree of conflict changes the path of this decision trajectory. Thus, the trajectory through the manifold of decision policies is solely sensitive to the presence of a change point rather than either the history of change point frequency or the history of optimal choice ambiguity. This suggests that adaptation to environmental changes in action-outcome contingencies involves a rapid, coordinated decrease in both the amount of information needed to make a decision and the rate of information accumulation, with a stereotyped return to a stable baseline soon thereafter until another change occurs.

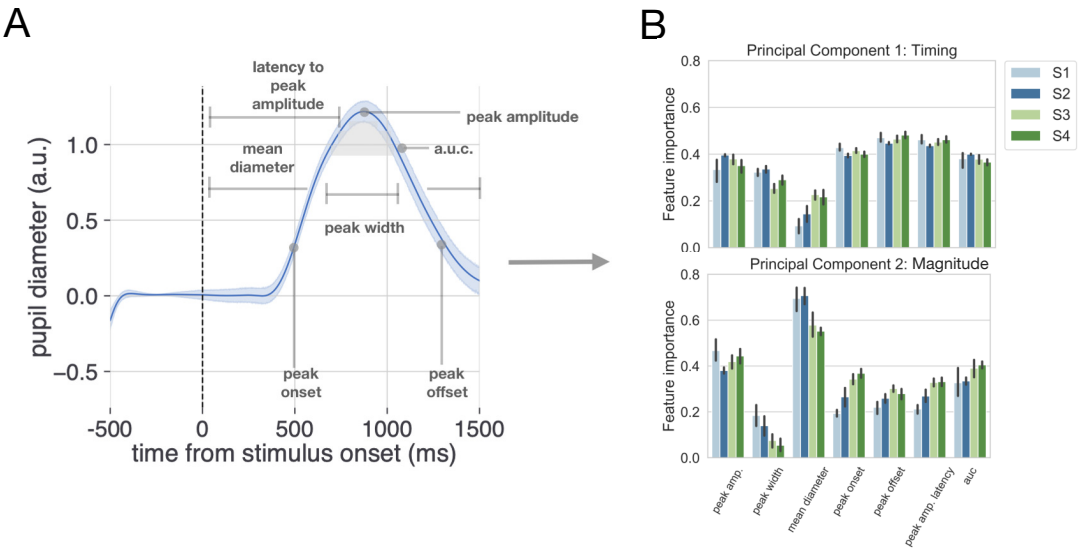


Figure 7. Method for analyzing pupil data. A) The evoked pupillary response was characterized according to seven metrics. B) These pupillary features were submitted to a principal component analysis. The contribution of each feature to the variance explained for the first two components is plotted for each subject.

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

The locus coereleus-norepinephrine (LC-NE) system is known to modulate exploration states under uncertainty and pupil diameter shows a tight correspondence with LC neuron firing rate (***Aston-Jones and Cohen, 2005***), with changes in pupil diameter indexing the explore-exploit decision state (***Jepma and Nieuwenhuis, 2011***). Similar to the classic Yerkes-Dodson curve relating arousal to performance (***Yerkes et al., 1908***), performance is optimal when tonic LC activity is moderate and phasic LC activity increases following a goal-related stimulus (***Aston-Jones et al., 1999***). 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 moderate a spike in phasic activity following a change in action-outcome contingencies.

We characterized the evoked pupillary response on each trial 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. 7A). We then submitted these metrics to principal component analysis to reduce their dimensionality while capturing maximum variance. Evoked response characterization and principal component analyses 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 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. 7). 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. 7B, upper panel). Mean pupil diameter and peak amplitude had the greatest feature importance for the second principal component (Fig. 7B, lower panel). Thus, for interpretability, we refer to the first and second principal components as timing and magnitude

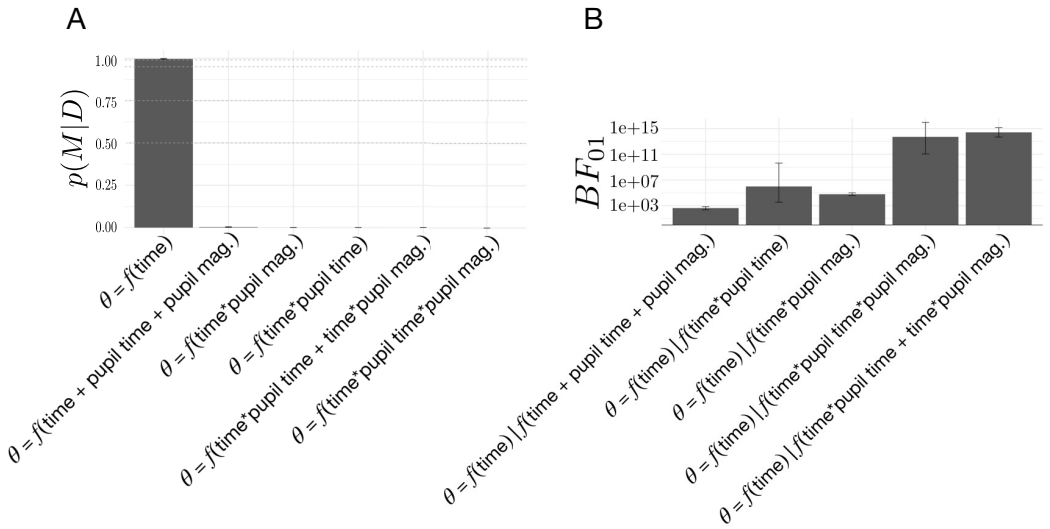


Figure 8. Model comparisons for the effect of pupillary dynamics on the relationship between drift rate and boundary height (θ). A) The posterior probability for models testing for an effect of pupillary dynamics on θ. B) The Bayes Factor for the time null model relative to the alternative models specifying an effect of pupillary dynamics on θ. Note that time refers to time relative to the onset of a change point. All models specifying an interaction also include main effects.

components, respectively (Fig. 7B).

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 pupillary response and shifts in θ. As in our previous model comparison (Fig. 6; see Environmental instability prompts a stereotyped decision trajectory), we found that the model describing the relationship between decision dynamics (θ) and time relative to a change point had the highest posterior probability given the set of models considered (Fig. 8A). To further evaluate the extent of the evidence for the time-null hypothesis, we present the evidence for the time null model against the original model set as BF_{01} (Fig. 8B). We find unambiguous evidence in favor of the time-null model relative to the models specifying the modulation of θ via the timing and magnitude features of the 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.

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 fast exploratory strategy by reducing both the drift rate and 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 highlight how the underlying decision algorithm adapts to different forms of uncertainty.

Exploration and exploitation states are not discrete, but exist along a continuum (***Addicott et al., 2017***). Instead of switching between binary states, humans manage environmental instability by adjusting the degree of exploration and exploitation (***Sadeghiyeh et al., 2020***; ***Prat-Carrabin et al., 2020***; ***Feng et al., 2020***; ***Wilson et al., 2014***; ***Payzan-LeNestour and Bossaerts, 2011***, ***2012***; ***Wilson***

et al., 2020). Depending on the relative configuration of parameters in the accumulation to bound process, this adjustment can manifest as either speeded or slowed decisions (Fig. 1E) (***Alexandrowicz, 2020***; ***Ratcliff, 1978***). Our results suggest that, in the context of volatile two-choice decisions, humans adopt a mechanism that simultaneously changes the information threshold and rate of evidence accumulation to adapt to an environmental change (Fig. 5A). 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. 5A). The rapid increase in change point probability causes a rapid, but slight, *contraction* of the boundary height on the subsequent trial, thereby lowering the criterion for selecting a new action and allowing variability in the accumulation process to have a greater influence on choice (Fig. 5B-C). These changes lead to fast exploratory decisions that facilitate discovery of the new optimal action and result in a quick recovery of the original threshold value. 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 lowering 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 quickly gathering information in an unbiased manner.

Critically, our finding that underlying decision policies can express exploration as a relative reconfiguration of multiple underlying decision parameters closely parallels recent work in the domain of information-seeking. Information seeking has been decomposed into random and directed components (***Wilson et al., 2014***). 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 (***Feng et al., 2020***). 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 (***Dunovan and Verstynen, 2019***; ***Urai et al., 2018***). Previous work by our lab (***Dunovan and Verstynen, 2019***) 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 (***Urai et al., 2018***). Our current findings and these previous observations (***Pedersen et al., 2017***; ***Ratcliff and Frank, 2012***) all highlight how sensitive the parameters of accumulation-to-bound processes are to immediate experience.

At first glance, our results might appear to conflict with prior findings linking relative differences in action value to an *increase* in boundary height (***Cavanagh et al., 2011***; ***Frank, 2006***; ***Frank et al., 2015***). However, in contrast with the previous studies, our experiments were designed to not only include an estimate of reward-conflict, but also an estimate of change point probability. This allowed us to directly compare models capturing the effect of conflict and volatility, respectively, on the drift rate and the boundary height (Table: Fig. 1). Including both of these estimates may parse drift rate effects from those on the boundary height, which would otherwise be obscured. Importantly, the model that mapped change point probability to the drift rate and belief to the boundary

height did not explain variability in choice and reaction times as well as the model in which change point probability drives changes in the boundary height and belief drives changes in the drift rate. This comparison lends some credence to the possibility that previous observations of a selective effect of conflict on the boundary height (***Cavanagh et al., 2011***; ***Frank, 2006***; ***Frank et al., 2015***) may have been the result of an obscured relationship between conflict and volatility.

We hypothesized that these shifts in decision policies would be linked to changes in phasic responses of the LC-NE pathways (***Aston-Jones and Cohen, 2005***). However, our results failed to find such an association. This suggests that an alternative mechanism drives the adaptive changes that we observed behaviorally. On possible alternative is the cortico-basal ganglia-thalamic (CBGT) pathways or "loops". Both recent experimental (***Yartsev et al., 2018***) and theoretical (***Bogacz and Larsen, 2011***; ***Caballero et al., 2018***; ***Wei et al., 2015***) 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 being ideal for implementing the sequential probability ratio test (***Bogacz and Gurney, 2007***; ***Bogacz, 2007***), the statistically optimal algorithm for evidence accumulation decisions and the basis for the DDM itself (***Ratcliff, 1978***). 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 (***Mikhael and Bogacz, 2016***; ***Bariselli et al., 2018***; ***Dunovan et al., 2019***; ***Rubin et al., 2020***), while the indirect pathways are linked to control of the decision threshold (***Wei et al., 2015***; ***Herz et al., 2016b***; ***Bogacz, 2007***; ***Ratcliff and Frank, 2012***). 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, these pathways are also a target of the dopaminergic signaling that drives reinforcement learning (***Schultz et al., 1992***), 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 (***Kakade and Dayan, 2002***) and recent pharmacological evidence suggests a role for dopaminergic regulation of exploration in humans (***Chakroun et al., 2020***). More explicitly, both directed and random exploration have been linked to variations in genes that affect dopamine levels in prefrontal cortex and striatum, respectively (***Gershman and Tzovaras, 2018***). 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 (***Vich et al., 2020***), with overall indirect pathway activity 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 (***Dunovan et al., 2019***; ***Rubin et al., 2020***). 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 (***Heston et al., 2020***). 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 ***Dunovan et al.*** (***2019***), and change-point-probability-driven changes to the boundary height varying with overall indirect pathway activity (***Dunovan et al., 2019***; ***Vich et al., 2020***). Future physiological studies should focus on validating this predicted relationship between decision policy reconfiguration and CBGT pathways.

Indeed, 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 (***Ratcliff, 1978***). 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. ***Tajima et al. (2019)***). Finally, because our estimate of the relationship between our ideal observer estimates of uncertainty and human estimates of uncertainty was indirect, this work would benefit from online approximations of ideal observer estimates, as has been done previously (***Wilson et al., 2010***). Indeed, there can be substantive individual differences in the detection of of change points (***Wilson et al., 2010***). 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.

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 less evidence is 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 fast exploratory state. As the environment becomes stable, the system slowly 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.

Methods and Materials

Participants

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. All participants were sourced from the Carnegie Mellon University community. Both experiments were approved by the Institutional Review Board at Carnegie Mellon University.

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 (***Foster and Deardorff, 2017***). It should be noted that the direction of the association between change point probability and boundary height was inverted in the preregistration.

Stimuli and Procedure

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. 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 distribution 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 (***Byrne et al., 2017***; ***Murray et al., 2009***), the order of conditions was counterbalanced across participants.

Experiment 2

Experiment 2 used male and female Greebles (***Gauthier and Tarr, 1997***) as selection targets (Fig. 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 (***Gauthier and Tarr, 1997***), with the goal of selecting the sex identity of the Greeble that was most profitable (Fig. 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 (***Hurwicz, 1972***; ***Ledyard, 1989***), 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, ≥ 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. 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 (***Sirois and Brisson, 2014***), 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.

Models and simulations

Q-Learning simulations

A simple, tabular q-learning agent (***Sutton and Barto, 1998***) 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

$$\pi_t = \frac{\exp^{\beta * Q_t}}{\Sigma \beta * Q_t}. \tag{1}$$

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

$$Q_{t+1} = Q_t + \alpha(reward - Q_t), \tag{2}$$

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.

Drift diffusion model simulations

A normative drift-diffusion model (DDM) process (***Ratcliff, 1978***) 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.

$$\begin{aligned} d\theta &= \upsilon dt + \sigma dW \quad \text{for} \quad t > tr; \\ \theta(t \leq tr) &= z/a \end{aligned} \tag{3}$$

where υ 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, υ reflects the signal-to-noise ratio of the stimulus. However, in a value-based decision task, υ 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 υ, 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 υ. 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 υ 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.

Cognitive model

Our *a priori* hypothesis was that the drift rate (υ) and the boundary height (a) should change on a trial-by-trial basis according to two estimates of uncertainty from an ideal observer (***Bond et al., 2018***). We adapted the below ideal observer calculations from a previous study (***Vaghi et al., 2017***) (for the original formulation of this reduced ideal observer model and its derivation, see (***Nassar et al., 2010***)).

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:

$$\begin{aligned} B_{t+1,c} &= B_{t,c} + \alpha_t \delta_t, \\ B_{t+1,u} &= B_{t,u}(1-\Omega_t) + \Omega_t E(r), \end{aligned} \tag{4}$$

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:

$$E(r) = \frac{\bar{r}_{t_0} + \bar{r}_{t_1}}{2}, \tag{5}$$

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:

$$\delta_t = r_t - B_{t,c}. \tag{6}$$

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:

$$\Delta B_{t+1} = B_{t,opt} - B_{t,subopt}. \tag{7}$$

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):

$$RU_t = \frac{\Omega_t \sigma_n^2 + (1-\Omega_t)(1-\phi_t)\sigma_n^2 + \Omega_t(1-\Omega_t)(\delta_t \phi_t)^2}{\Omega_t \sigma_n^2 + (1-\Omega_t)(1-\phi_t)\sigma_n^2 + \Omega_t(1-\Omega_t)(\delta_t \phi_t)^2 + \sigma_n^2}. \tag{8}$$

Thus ϕ is calculated as:

$$\phi_{t+1} = 1 - RU_t. \tag{9}$$

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

$$\sigma_t^2 = \sigma_n^2 + \frac{(1-\phi_t)\sigma_n^2}{\phi_t} \tag{10}$$

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):

$$\alpha_t = \Omega_t + (1-\Omega_t)(1-\phi_t). \tag{11}$$

To model how learners update action-values, we calculated an estimate of how often the same action gave a different reward (***Vaghi et al., 2017***). 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:

$$\Omega_t = \frac{U(r_t)H}{U(r_t)H + N(r_t | B_{\Delta_t}, \sigma_t^2)(1-H)}. \tag{12}$$

In equation (12), H refers to the hazard rate, or the global probability of a change point over trials:

$$H = \frac{n_{cp}}{n_{trials}}. \tag{13}$$

We expected 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:

$$v_{t+1} = \hat{\beta}_v \cdot \Delta B_t + v_t \tag{14}$$

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

$$a_{t+1} = a_0 - \hat{\beta}_a \cdot \Omega_t. \tag{15}$$

Hierarchical 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) updated the rate of evidence accumulation (v) and the amount of evidence needed to make a decision (a), we regressed the ideal observer estimates onto the decision parameters using hierarchical drift diffusion model (HDDM) regression (***Wiecki et al., 2013***). 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. 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.

Markov-chain Monte-Carlo methods were used to sample the posterior distributions of the regression coefficients. Ten thousand samples were drawn from the posterior distributions of the coefficients for each model. 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 identify the regression model that best accounted for the data, we conducted a model selection process using Deviance Information Criterion (DIC) scores. We compared the set of models (Table 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 (***Spiegelhalter et al., 2002***; ***Burnham and Anderson, 1998***).

To test our hypotheses regarding these HDDM regression estimates, we 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. Trial-by-trial estimates of boundary height and drift rate were calculated from the regression equation for the winning model.

For Experiment 2, a regression model as described above was fit to each subject.

Analyses

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 (Acknowledgments). 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 (***Satterthwaite, 1946***; ***Luke, 2017***).

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. 5B).

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. 1). 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 (***Mulder and Klugkist, 2017***) was performed to determine how θ changed as a function of the number of trials after the change point:

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

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. 6) for sequential pairs of trials had equal means ($\theta_{\Delta_t} = \theta_{\Delta_{t+1}}$). This result (Supp. Fig. 7) 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 pro-

posed (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 (***Kruschke and Vanpaemel, 2015***). 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.

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 (***Beatty, 1982***), 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. 8 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. 7A). 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. 7B).

Model proposals and evaluation

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

$$\theta = \beta_0.$$

The second, time-null model estimates decision dynamics solely as a function of time, with the timescale of consideration determined by the results of the stability convergence analysis from Experiment 1:

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

We first evaluated whether the posterior probability of the time-null 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 time-null model and the CIs were non-overlapping), we proceeded to evaluate the evidence for alternative models relative to this time-null 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 pupillary response on boundary height and drift rate dynamics (see Fig. 8 for the full set of models considered). The first two principal component projections 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 decision dynamics, we estimated 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 1 for the full set of models considered).

We used Bayes Factors to quantify the ratio of evidence for competing hypotheses (***Wagenmakers, 2007***). To estimate whether these models accounted for decision dynamics beyond the effect of time alone, we calculate the Bayes Factor for the time-null 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 (***Wagenmakers, 2007***). 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.

Acknowledgments

We thank all members of the Cognitive Axon Lab for their feedback during the development of this work. We also thank Marlene Behrmann and Michael Granovetter for their help with eye-tracking and pupillometry data collection.

References

Addicott MA, Pearson JM, Sweitzer MM, Barack DL, Platt ML. A primer on foraging and the explore/exploit trade-off for psychiatry research. Neuropsychopharmacology. 2017; 42(10):1931–1939.

Alexandrowicz RW. The diffusion model visualizer: an interactive tool to understand the diffusion model parameters. Psychological research. 2020; 84(4):1157–1165.

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

Aston-Jones G, Cohen JD. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci. 2005; 28:403–450.

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

Bariselli S, Fobbs W, Creed M, Kravitz A. A competitive model for striatal action selection. Brain research. 2018; .

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

Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. Nature neuroscience. 2007; 10(9):1214.

Bland AR, et al. Different varieties of uncertainty in human decision-making. Frontiers in neuroscience. 2012; 6:85.

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

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

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

Bogacz R, Wagenmakers EJ, Forstmann BU, Nieuwenhuis S. The neural basis of the speed–accuracy tradeoff. Trends in neurosciences. 2010; 33(1):10–16.

Bond K, Dunovan K, Verstynen TD, The influence of volatility and conflict on adaptive decision making. OSF; 2018. osf.io/5esn4.

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

Burnham KP, Anderson DR. Practical use of the information-theoretic approach. In: *Model selection and inference* Springer; 1998.p. 75–117.

Byrne JE, Hughes ME, Rossell SL, Johnson SL, Murray G. Time of day differences in neural reward functioning in healthy young men. Journal of Neuroscience. 2017; p. 0918–17.

Caballero JA, Humphries MD, Gurney KN. A probabilistic, distributed, recursive mechanism for decision-making in the brain. PLoS Comput Biol. 2018 Apr; 14(4):e1006033.

Cavanagh JF, Wiecki TV, Cohen MX, Figueroa CM, Samanta J, Sherman SJ, Frank MJ. Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. Nature neuroscience. 2011; 14(11):1462.

Chakroun K, Mathar D, Wiehler A, Ganzer F, Peters J. Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. Elife. 2020; 9:e51260.

Daw ND, O'doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441(7095):876.

Dayan P, Yu AJ. Phasic norepinephrine: a neural interrupt signal for unexpected events. Network: Computation in Neural Systems. 2006; 17(4):335–350.

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

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

Feng SF, Wang S, Zarnescu S, Wilson RC, The dynamics of explore-exploit decisions reveal a signal-to-noise mechanism for random exploration. PsyArXiv; 2020. psyarxiv.com/uepr7, doi: 10.31234/osf.io/uepr7.

Forstmann BU, Anwander A, Schäfer A, Neumann J, Brown S, Wagenmakers EJ, Bogacz R, Turner R. Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. Proc Natl Acad Sci U S A. 2010 Sep; 107(36):15916–15920.

Forstmann BU, Dutilh G, Brown S, Neumann J, von Cramon DY, Ridderinkhof KR, Wagenmakers EJ. Striatum and pre-SMA facilitate decision-making under time pressure. Proc Natl Acad Sci U S A. 2008; 105(45):17538–17542.

Foster ED, Deardorff A. Open science framework (OSF). Journal of the Medical Library Association: JMLA. 2017; 105(2):203.

Frank MJ. Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Networks. 2006; 19(8):1120–1136.

Frank MJ, Gagne C, Nyhus E, Masters S, Wiecki TV, Cavanagh JF, Badre D. fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. Journal of Neuroscience. 2015; 35(2):485–494.

Gauthier I, Tarr MJ. Becoming a "Greeble" expert: Exploring mechanisms for face recognition. Vision research. 1997; 37(12):1673–1682.

Gershman SJ, Tzovaras BG. Dopaminergic genes are associated with both directed and random exploration. Neuropsychologia. 2018; 120:97–104.

Herz DM, Tan H, Brittain JS, Fischer P, Cheeran B, Green AL, FitzGerald J, Aziz TZ, Ashkan K, Little S, Foltynie T, Limousin P, Zrinzo L, Bogacz R, Brown P. Distinct mechanisms mediate speed-accuracy adjustments in cortico-subthalamic networks. Elife. 2017 Jan; 6.

Herz DM, Zavala BA, Bogacz R, Brown P. Neural correlates of decision thresholds in the human subthalamic nucleus. Current Biology. 2016; 26(7):916–920.

Herz DM, Zavala BA, Bogacz R, Brown P. Neural correlates of decision thresholds in the human subthalamic nucleus. Current Biology. 2016; 26(7):916–920.

Heston J, Friedman A, Baqai M, Bavafa N, Aron AR, Hnasko TS. Activation of subthalamic nucleus stop circuit disrupts cognitive performance. eNeuro. 2020; .

Hurwicz L. On Informationally Decentralized Systems. Decision and Organization. 1972; p. 320.

Jeffreys H. The theory of probability. OUP Oxford; 1998.

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

Kakade S, Dayan P. Dopamine: generalization and bonuses. Neural Networks. 2002; 15(4-6):549–559.

Kruschke JK, Vanpaemel W. Bayesian estimation in hierarchical models. The Oxford handbook of computational and mathematical psychology. 2015; p. 279–299.

Ledyard JO. Incentive compatibility. In: *Allocation, Information and Markets* Springer; 1989.p. 141–151.

Luke SG. Evaluating significance in linear mixed-effects models in R. Behavior research methods. 2017; 49(4):1494–1502.

Mendonça AG, Drugowitsch J, Vicente MI, DeWitt EE, Pouget A, Mainen ZF. The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. Nature Communications. 2020; 11(1):1–15.

Mikhael JG, Bogacz R. Learning Reward Uncertainty in the Basal Ganglia. PLoS Comput Biol. 2016 Sep; 12(9):e1005062.

Mulder K, Klugkist I. Bayesian estimation and hypothesis tests for a circular generalized linear model. Journal of mathematical psychology. 2017; 80:4–14.

Murray G, Nicholas CL, Kleiman J, Dwyer R, Carrington MJ, Allen NB, Trinder J. Nature's clocks and human mood: The circadian system modulates reward motivation. Emotion. 2009; 9(5):705.

Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI. Rational regulation of learning dynamics by pupil-linked arousal systems. Nature neuroscience. 2012; 15(7):1040.

Nassar MR, Wilson RC, Heasly B, Gold JI. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. Journal of Neuroscience. 2010; 30(37):12366–12378.

O'reilly JX. Making predictions in a changing world—inference, uncertainty, and learning. Frontiers in neuroscience. 2013; 7:105.

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

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

Pedersen ML, Frank MJ, Biele G. The drift diffusion model as the choice rule in reinforcement learning. Psychonomic bulletin & review. 2017; 24(4):1234–1251.

Prat-Carrabin A, Wilson RC, Cohen JD, Da Silveira RA. Human Inference in Changing Environments with Temporal Structure. BioRxiv. 2020; p. 720516.

Ratcliff R. A theory of Memory Retrival. Psychol Rev. 1978; 85(2):59–108.

Ratcliff R, Frank MJ. Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. Neural computation. 2012; 24(5):1186–1229.

Rubin JE, Vich C, Clapp M, Noneman K, Verstynen T. The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution. European Journal of Neuroscience. 2020; .

Sadeghiyeh H, Wang S, Alberhasky MR, Kyllo HM, Shenhav A, Wilson RC. Temporal discounting correlates with directed exploration but not with random exploration. Scientific reports. 2020; 10(1):1–10.

Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics bulletin. 1946; 2(6):110–114.

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

Sirois S, Brisson J. Pupillometry. Wiley Interdisciplinary Reviews: Cognitive Science. 2014; 5(6):679–692.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. Journal of the royal statistical society: Series b (statistical methodology). 2002; 64(4):583–639.

Sutton RS, Barto AG. Introduction to Reinforcement Learning. Camgridge: MIT Press; 1998.

Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.

Tajima S, Drugowitsch J, Patel N, Pouget A. Optimal policy for multi-alternative decisions. Nature neuroscience. 2019; 22(9):1503–1511.

Urai AE, Braun A, Donner TH. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. Nature communications. 2017; 8(1):1–11.

Urai AE, De Gee JW, Donner TH. Choice history biases subsequent evidence accumulation. BioRxiv. 2018; p. 251595.

Vaghi MM, Luyckx F, Sule A, Fineberg NA, Robbins TW, De Martino B. Compulsivity reveals a novel dissociation between action and confidence. Neuron. 2017; 96(2):348–354.

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

Wagenmakers EJ. A practical solution to the pervasive problems ofp values. Psychonomic bulletin & review. 2007; 14(5):779–804.

Wei W, Rubin JE, Wang XJ. Role of the indirect pathway of the basal ganglia in perceptual decision making. Journal of Neuroscience. 2015; 35(9):4052–4064.

Wiecki TV, Sofer I, Frank MJ. HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. Frontiers in neuroinformatics. 2013; 7:14.

Wilson RC, Bonawitz E, Costa VD, Ebitz RB. Balancing exploration and exploitation with information and randomization. Current Opinion in Behavioral Sciences. 2020; 38:49–56.

Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD. Humans use directed and random exploration to solve the explore–exploit dilemma. Journal of Experimental Psychology: General. 2014; 143(6):2074.

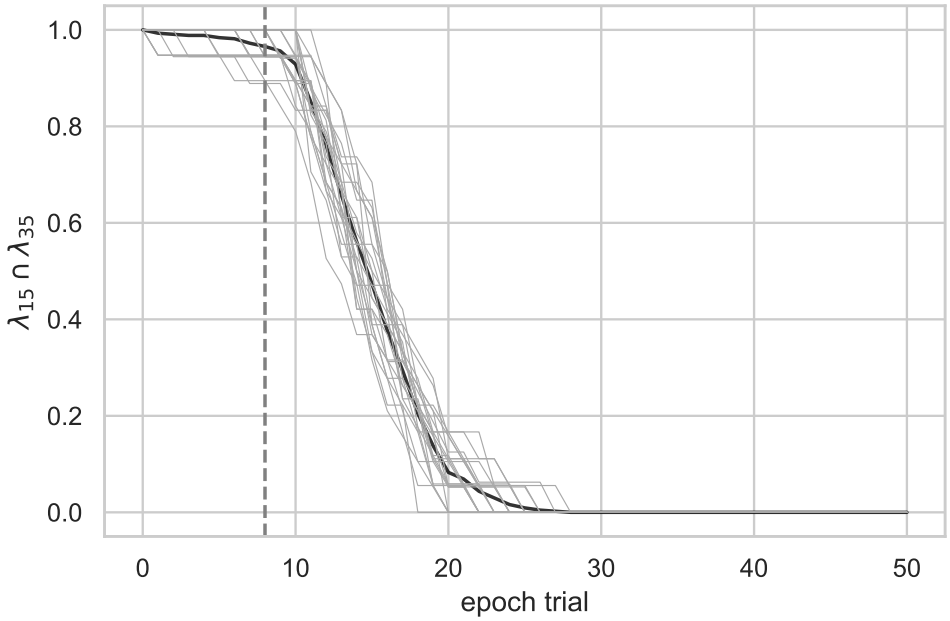
Wilson RC, Nassar MR, Gold JI. Bayesian online learning of the hazard rate in change-point problems. Neural computation. 2010; 22(9):2452–2476.

Wilson RC, Niv Y. Inferring relevance in a changing world. Frontiers in human neuroscience. 2012; 5:189.

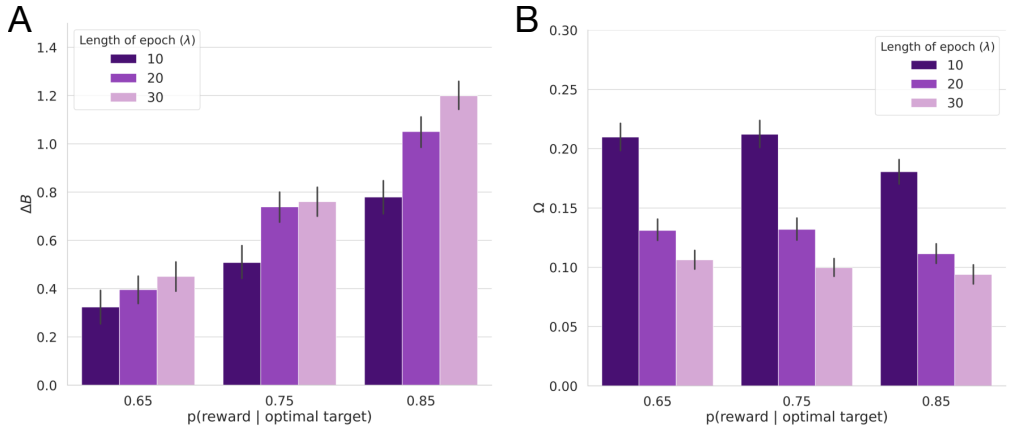
Yartsev MM, Hanks TD, Yoon AM, Brody CD. Causal contribution and dynamical encoding in the striatum during evidence accumulation. Elife. 2018 Aug; 7:e34929.

Yerkes RM, Dodson JD, et al. The relation of strength of stimulus to rapidity of habit-formation. Punishment: Issues and experiments. 1908; p. 27–41.

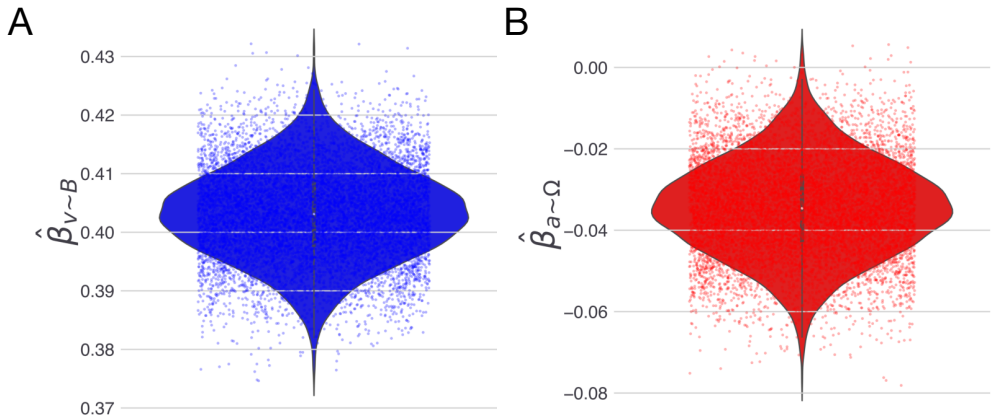
Appendix 1



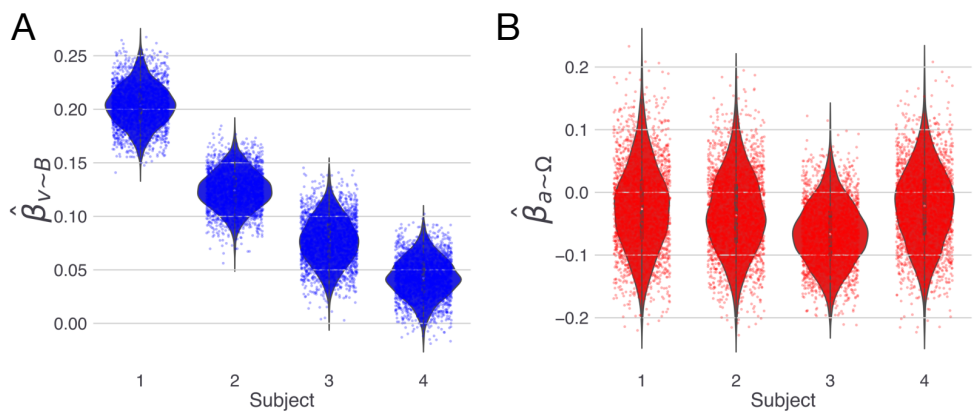
Appendix 1 Figure 1. Initial window selection.
Analysis conducted on data from Experiment 1 to determine the timescale of the response that maximized the intersection between high volatility ($\lambda = 15$) and low volatility ($\lambda = 35$) data. The bolded line represents the mean and the gray lines represent individual subjects. The dotted line indicates the initial window of nine trials used.



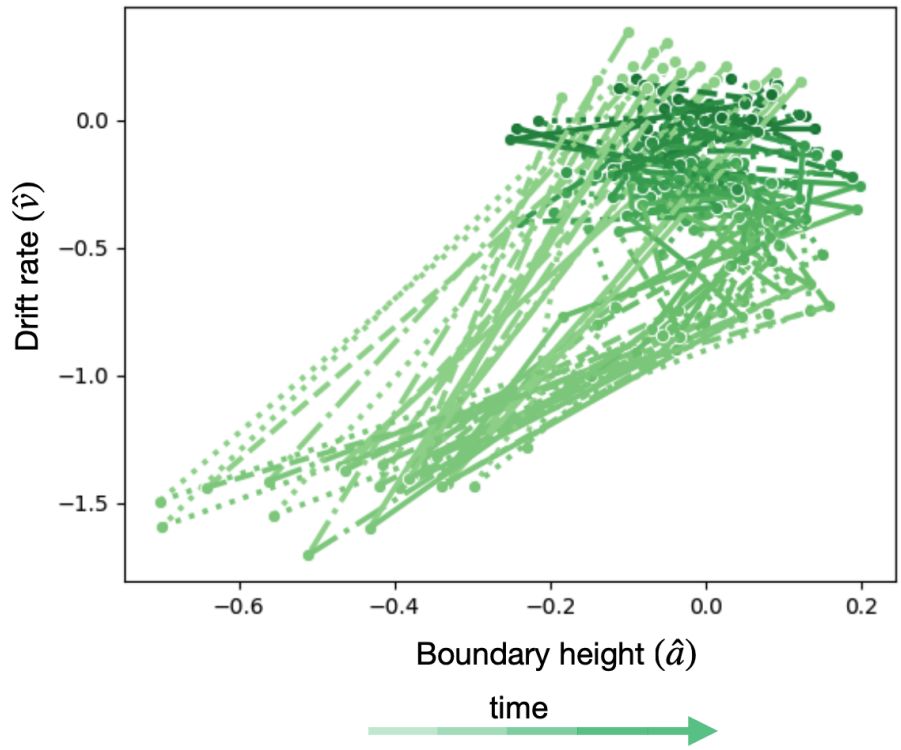
Appendix 1 Figure 2. Ideal observer estimates for Experiment 2.
A) The average belief in the value of the optimal target (ΔB) as a function of the probability of reward (conflict) and the average period of stability for the optimal choice (λ; volatility). B) Average change point probability (Ω) as a function of conflict and volatility.



Appendix 1 Figure 3. Distributions of HDDM β coefficients for Experiment 1.
A) The distribution of the estimate of the relationship between the drift rate (v) and the belief in the value of the optimal choice (ΔB). B) The distribution of the estimate of the relationship between the boundary height (a) and change point probability (Ω).

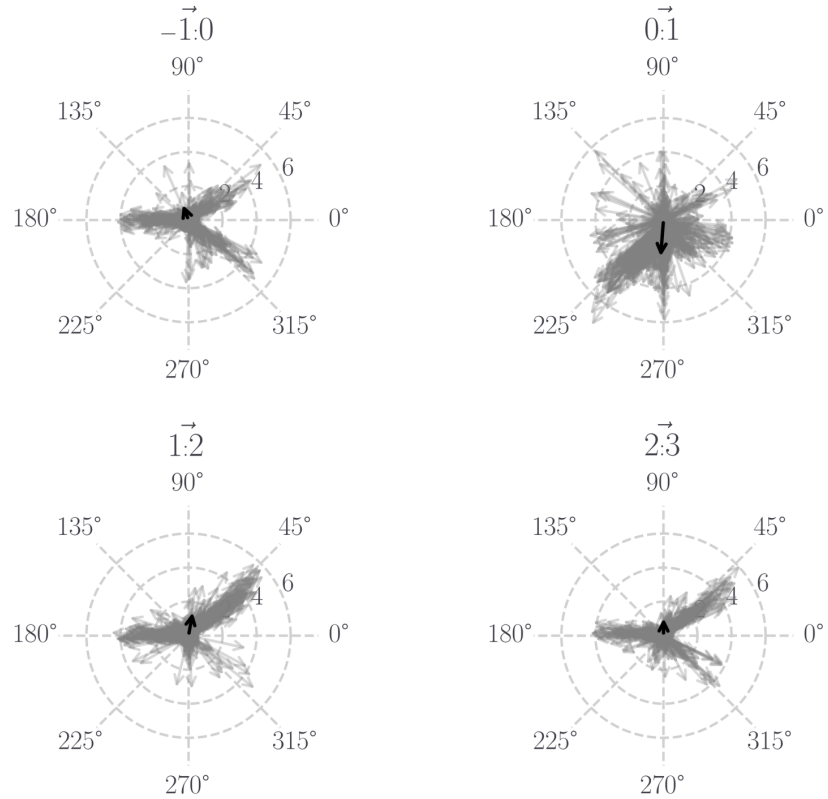


Appendix 1 Figure 4. Distributions of HDDM β coefficients for Experiment 2.
A) The relationship between the drift rate (v) and the belief in the value of the optimal choice (ΔB). B) The relationship between the boundary height (a) and change point probability (Ω).



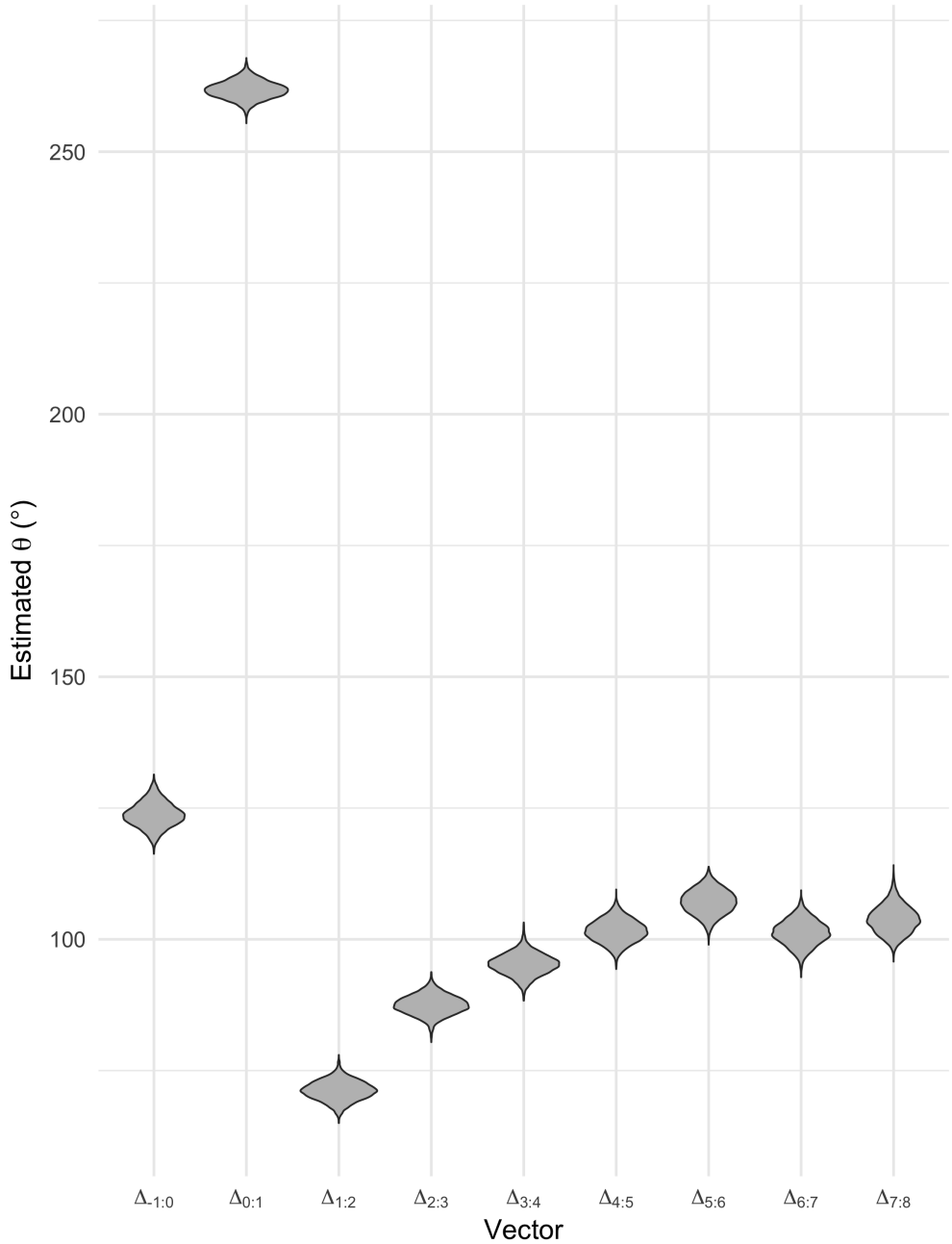
Appendix 1 Figure 5. Decision trajectories for Experiment 1.

The change in boundary height and drift rate over time for Experiment 1. Each line represents the data for one participant.



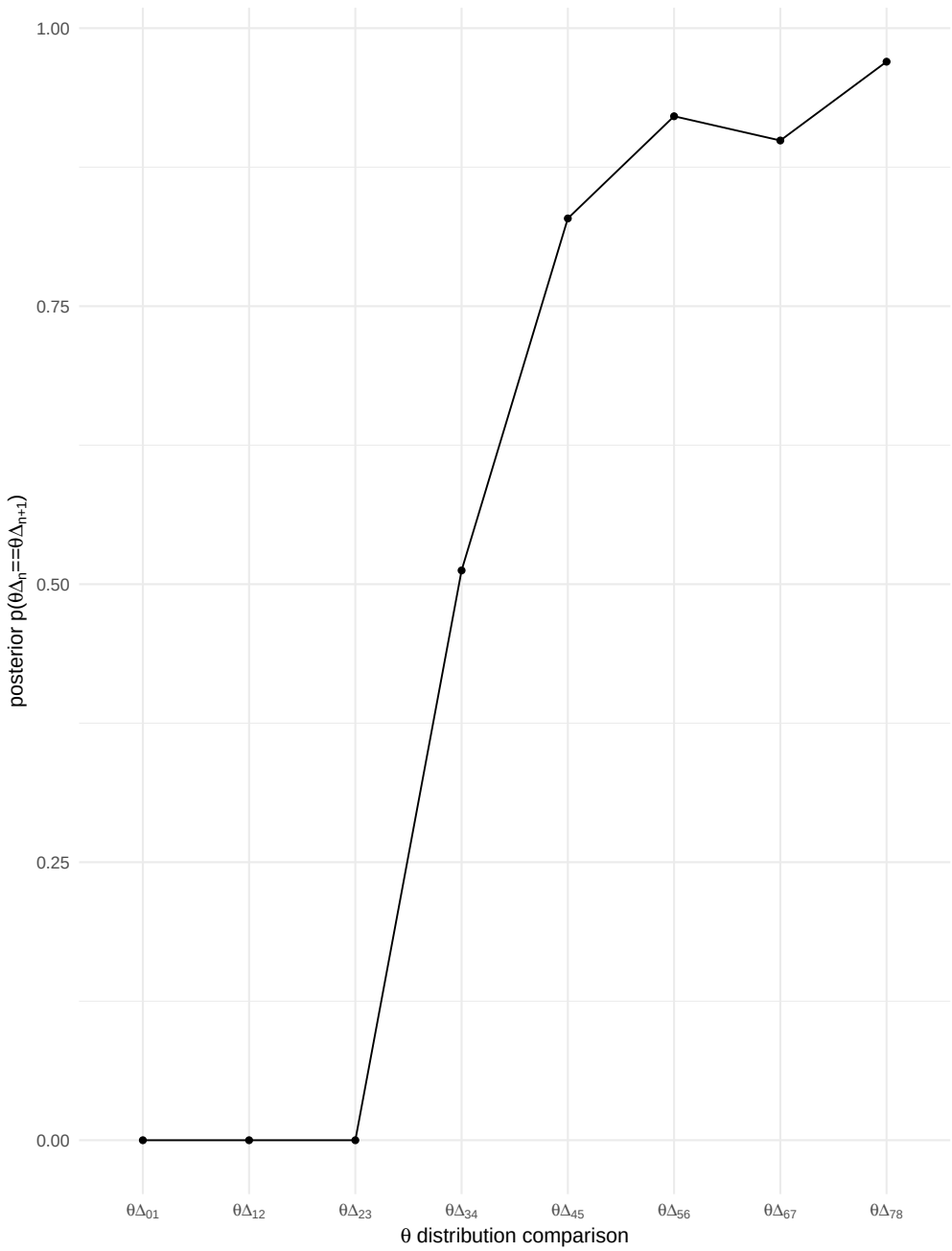
Appendix 1 Figure 6. Vector distributions.

The distribution of vectors formed between sequential (a, v) coordinates, beginning with the trial prior to the change point. The black arrow shown on each panel represents the mean vector.



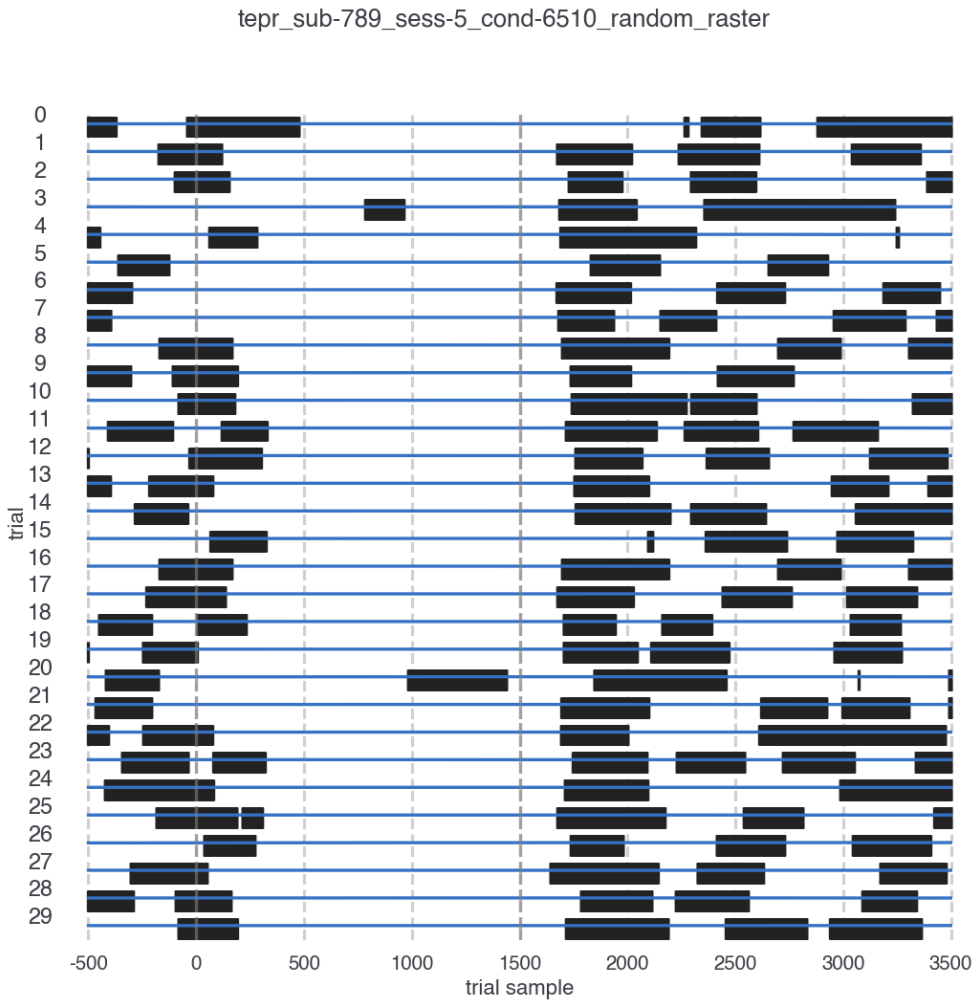
Appendix 1 Figure 7. Stability analysis.

Analysis conducted on Experiment 1 to determine the timescale of the response to consider for Experiment 2. The estimated angle is plotted as a function of time within an epoch (estimate from circular regression).



Appendix 1 Figure 8. Quantification of stability.

Probability that sequential posterior distributions for $\beta_{\Delta t}$ have equal means.



Appendix 1 Figure 9. Blink timing.

Blink timing for a sample participant. For visibility, thirty trials were selected at random. The onset of the trial is marked as time 0 and the trial ends at 1500 ms. Blinks are marked in black. Blink timing plots are available for all subjects and all conditions in the GitHub repository for this publication.

Appendix 2

	B	Ω	DIC	ΔDIC_{null}	ΔDIC_{best}
*I	v	a	-101886.4	-15477.5	0.0
II	a	v	-87486.7	-1077.8	14399.7
III	–	v	-87373.9	-965.0	14512.5
IV	v	–	-97634.70	-11225.8	4251.70
V	–	a	-90577.3	-4168.40	11309.00
VI	a	–	-86525.7	-116.70	15360.7
VII	–	–	-86408.9	0.0	15477.5

Appendix 2 Table 1. Power analysis for Experiment 1.

The results of the model comparison analysis using simulated data. 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 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 last row represents the null, intercept-only regression model. The best performing model is denoted by an asterisk.