When to choose: The role of information seeking in the speed-accuracy tradeoff

Anonymous Author(s)
Affiliation
Address
email

Abstract

Normative accounts of decision-making predict that people attempt to balance the immediate reward associated with a correct response against the cost of deliberation. However, humans frequently deliberate longer than normative models say they should. We propose that people try to optimize not only their rate of material rewards, but also their rate of information gain. A computational model that combines this idea with a standard drift diffusion process reveals that an agent programmed to maximize a combination of reward and information rates acts like human decision makers, reproducing key patterns of behavior not predicted by existing models. Moreover, if we assume that skill level is sensitive to deliberation time, a novice agent who maximizes even a small amount of information rate will often earn more reward in the long run than one who only maximizes reward rate. Maximizing a combination of reward and information rate is a relatively simple and myopic strategy, but approximates optimal behavior over learning, making it a candidate heuristic for this difficult intertemporal choice problem.

1 Introduction

Motivation. Normative accounts of decision-making predict that humans attempt to maximize the rate at which they are rewarded by optimizing how long they spend deliberating: long enough to make informed decisions, but not so long as to waste precious time [7, 8]. Mysteriously, this prediction is often violated. Humans systematically fail to maximize their reward rate by spending too much time deliberating [4, 1]. Several accounts of why humans fail to maximize reward rate have been offered [3, 19], but none provide a particularly good fit to the data. We propose that the systematic failure of humans to select decision times that maximize reward rate reflects the intrinsic value that humans assign to information [15, 2]. When determining how long to deliberate, people try to optimize not only their rate of reward, but also their rate of information gain.

Background. In a standard drift-diffusion model (DDM), decision-makers accumulate noisy evidence over time [14]. Given a set of prior beliefs about the evidence-generating process, decision-makers can use the accumulated evidence to compute, at each moment in time, the posterior probability that a given option is correct [7, 13, 6]. A decision is made once the accumulated evidence reaches a predetermined bound [14], which denotes how confident a decision-maker needs to be about which option is correct before making a decision. This bound can be constant or change over time [7, 13], and its location will impact the agent's reward rate. There exists an optimal bound for every signal-to-noise ratio (SNR) and inter-trial interval that can be parametrized as a relation between decision time (DT) and error rate (ER) called the optimal performance curve (OPC) [3].

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

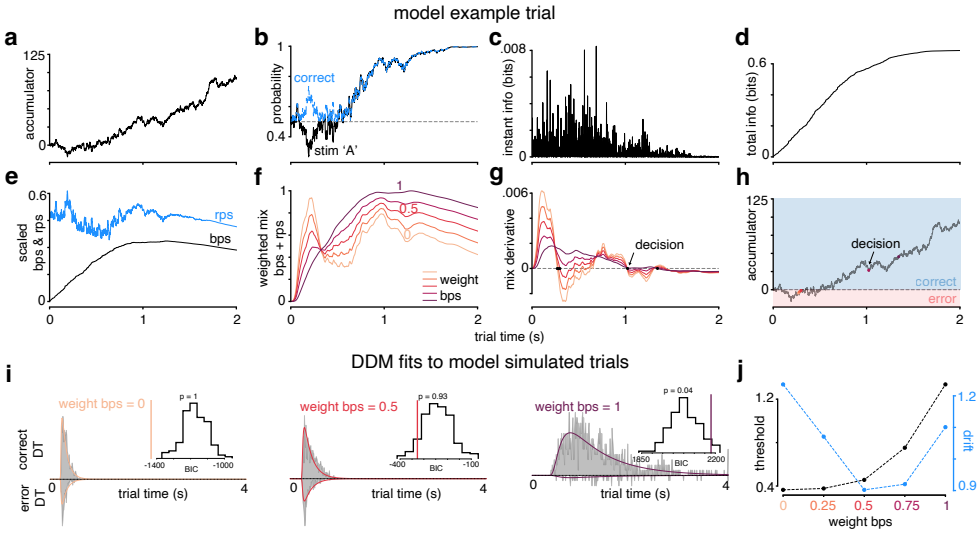

Figure 1: **Model description and decision time analysis.** **(a)** accumulator state. **(b)** p(stimulus 'A') and p(correct) (x, confidence). **(c)** instantaneous info i_n via eq. 1. **(d)** total info (sum over i_n). **(e)** within-trial reward rate $\hat{r}_n$ (rps) and information rate $\hat{i}_n$ (bps) via eqs. 2,3. **(f)** scaled (0 to 1) bps and rps are combined via a weighted sum where w (weight bps) controls the weight of bps relative to rps. This reward-information rate mix $\hat{m}_n$ is passed through a Kalman filter for smoothing. **(g)** a decision is made when $\hat{m}_n$ has peaked ($d\hat{m}_n/dt = 0$). **(h)** accumulator state at DT determines the outcome. **(i)** DDM fits to model simulations. Insets: DDM data generated with the fitted parameters and re-fit. p-value: simulation BICs larger (worse) than model's BIC. **(j)** Fitted DDM parameters by w.

2 Model & Results

Model overview. We propose a DDM whereby an agent makes a decision when a mixture of within-trial reward rate and information rate is maximized. Both rates depend on the posterior probability of a correct response at each moment of an accumulation process, which can be computed in closed form [7, 13, 6]. Within-trial reward rate corresponds to the probability of a correct response over time; within-trial information rate corresponds to the total amount of information gained over time.

Model description. An agent faced with a binary forced-choice task accumulates evidence according to a standard DDM [14] (Fig.1a). Assuming a Bayesian observer, the accumulator state (i.e., the total evidence accumulated) is used to compute, at every time step, the probability that each of the two response options is correct [7, 13, 6] (Fig. 1b). These probabilities are used to compute "confidence," denoted as $x \in [.5, 1]$: the probability of a correct decision assuming a "greedy" response strategy of always choosing the most probable option. Hence, if q and $1 - q$ are the probabilities that response "A" and "B" are correct, then $x = \max(q, 1 - q)$. We assume the agent has access to this confidence measure and uses it to estimate both reward rate and information rate within-trial.

Let i_n be the amount of information gained at time n,

$$i_n = D_{KL}[x_{0:n} || x_{0:n-1}] \tag{1}$$

$x_{0:n}$ denotes confidence conditional on all accumulator states observed from time 0 to time n, and $x_{0:n-1}$ denotes confidence conditional on all accumulator states observed from time 0 to time $n - 1$. D_{KL} is Kullback-Leibler divergence, which, in this case, quantifies the relative entropy from $x_{0:n-1}$ to $x_{0:n-1}$. In other words, i_n quantifies the amount of information that the decision-maker's nth accumulator state provides about her probability of a correct response (Fig.1c).

We assume that decision-makers track their rate of information gain at each time step n, denoted as $\hat{i}_n \in \mathbb{R}_{\geq 0}$, which equals the total information gained through the current time step (Fig.1d) divided by time spent accumulating n, non-decision time t_0, and the mean response-to-stimulus interval (RSI). We express it in units of bits-per-second (bps) (Fig.1e). We assume decision-makers similarly track their within-trial rate of reward $\hat{r}_n$, or rewards-per-second (rps), via their confidence (Fig.1e).

$$\text{bps} = \hat{i}_n = \frac{\sum_{j=0}^{n} i_j}{n + t_0 + \text{RSI}} \tag{2} \qquad \text{rps} = \hat{r}_n = \frac{x_{0:n}}{n + t_0 + \text{RSI}} \tag{3}$$

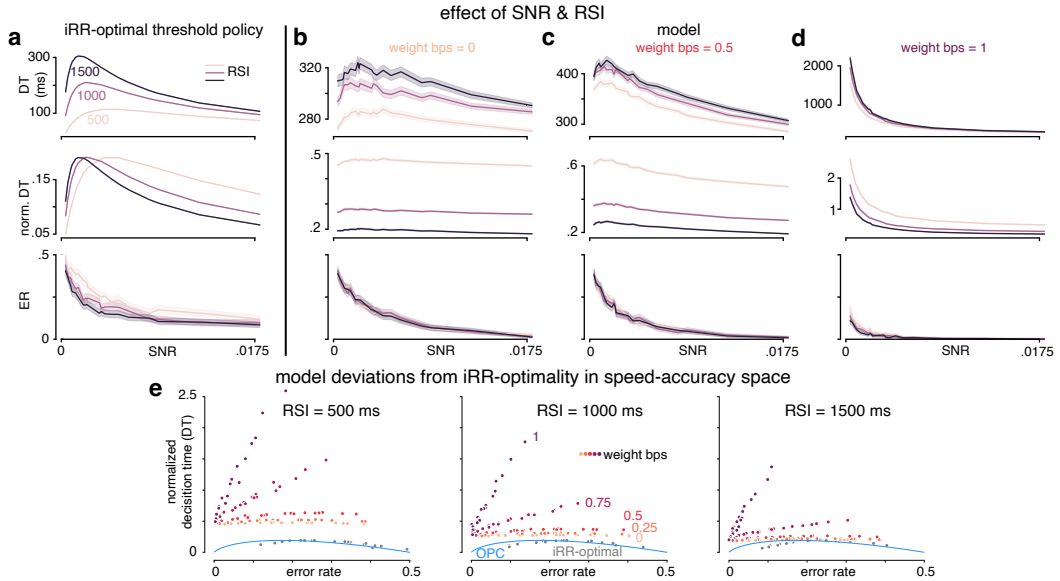

Figure 2: **Effects of SNR and RSI on speed and accuracy (a)** iRR-optimal threshold policy was used to simulate DT, normalized DT and ER over SNR. This policy predicts faster responses for low SNR. **(b,c,d)** model predicts relatively faster responses for low SNR as w increases. However, the range of DTs across RSI is lower because model is less sensitive to RSI. **(e)** ER & DT as functions of SNR (dots), RSI (columns), and w (warm tones) in speed-accuracy space. iRR-optimal policy in grey. OPC in blue.

According to our model, decision-makers attempt to maximize a mixture of reward rate and rate of information gain by stopping the accumulation process when this mixture, $\hat{m}_n$, is maximized (Fig.1f). The two quantities are first scaled separately so they are comparable, and then mixed through a weighted sum, $\hat{m}_n = w\hat{i}_n^{scaled} + (1-w)\hat{r}_n^{scaled}$, where w determines how much to prioritize information over reward, which could be plausibly learned as a prior over time. We assume the agent tracks the (smoothed) derivative of $\hat{m}_n$ via Kalman filtering, and makes a choice at the time n when the (smoothed) derivative is equal to zero, indicating that a local maximum has been reached (Fig.1g). The accumulator state at that time determines the outcome (Fig. 1h). We evaluate this model for different values of w. We also compare our results to those of an idealized policy that perfectly adheres to the OPC by maximizing instantaneous reward rate (i.e., an iRR-optimal policy).

Model produces realistic response time distributions. A hallmark of the standard DDM is that it predicts DT distributions that closely match behavior [14]. For different w, we fit 1000 simulated trials with a simple DDM, finding reasonable fits and plausible decision times (Fig. 1i). As expected, the DDM fits predicted increasing threshold with increasing w, and intriguingly a U-shaped function for drift, perhaps indicating accumulation is more challenging when split across objectives (Fig. 1j).

Empirical support. *Effects of SNR and RSI on speed and accuracy*. Among human and nonhuman decision-makers, speed and accuracy relate to signal-to-noise ratio (SNR: drift/noise) and inter-trial time (RSI) in ways that any viable model must reproduce. Figure 2a-d displays the independent effects of SNR and RSI on speed and accuracy for different w. As w increases, speed becomes an increasing function of SNR and a decreasing function of RSI. Notably, this is true neither for the iRR-optimal policy nor for $w = 0$, both of which predict faster DTs for low SNR, a violation of the standard finding that people respond more slowly as noise increases [4, 19, 1]. However, the DTs for $w = 0$ are still considerably larger than those from an iRR-optimal policy. For all mixture weights w, accuracy is an increasing function of SNR, consistent with empirical findings. We find, however, that the model is less sensitive to RSI than an iRR-optimal policy, evinced by the large difference in normalized DTs ($\mathrm{DT}/(t_0 + \mathrm{RSI})$) across RSIs. This lack of sensitivity could be due in part to the settings on the Kalman filter when calculating $\hat{m}_n$, preventing very-fast choices.

Deviations from optimality. Across all parameter settings, the model yields slower-than-optimal responses, in line with empirical findings in humans and nonhuman animals [4, 19, 1, 11]. In addition,

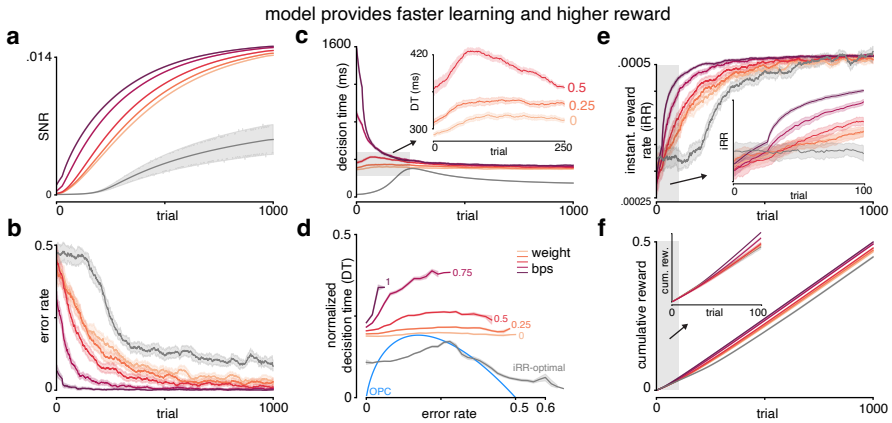

Figure 3: **Model provides faster learning and higher reward. (a,b,c,e,f)** SNR, ER, DT, iRR and cumulative reward over trials, separated by w (weight bps, warm tones) and iRR-optimal policy (grey). **(d)** trajectory in speed-accuracy space during learning. Insets: closeups of regions in grey.

overly slow performance is especially pronounced at low SNRs, and converges toward the OPC as SNR increases—a pattern also observed in the empirical data.

Normative support. *Faster learning and higher reward.* To measure the potential benefits of a decision policy that prioritizes information, we assume that skill level improves as a function of cumulative deliberation time (see Appendix). Our model leads to faster learning compared to an iRR-optimal agent equipped with the same learning capabilities (Fig. 3a,b). Notably, even the agent that only prioritizes reward ($w = 0$) learned faster than the iRR-optimal agent. This is because the model agent is maximizing reward rate within-trial and is beholden to noise, whereas the iRR-optimal agent maximizes reward rate across trials and sets an optimal threshold before each trial begins. Our model predicts slower responses at the beginning of learning, with the slowest responses at high w (Fig. 3c), and produces slower-than-optimal responses for high ERs, converging near the OPC for low ERs (Fig. 3d), in line with recent learning results in rats [11]. Although the model was suboptimal as judged by across-trial iRR, it led to a faster increase in iRR and a higher cumulative reward, lending normative support over an iRR-optimal agent (Fig. 3e,f).

3 Discussion

We have proposed an alternative decision strategy to a threshold during an evidence accumulation process based on the maximization of reward and information rates. Our model qualitatively recreates observed behavior in human and non-human subjects. Notably, our model predicts slow responses when noise is high. Current models miss this pattern on two levels. (1) Within-trials, a well-known limitation of the standard DDM [14, 10] and related models [18, 5, 17] is that increasing noise results in fast responses because the accumulator is more likely to randomly hit a response boundary. Our model eliminates this misprediction by introducing a novel dependency between noise and DT: as noise increases it takes longer for i_n to reach its maximum value. (2) Across-trials, normative models advocate for fast responses when noise is high [3], but behavioral responses fail to match this prediction.

Beyond empirical support, our model enjoys normative support when learning is taken into account, leading to faster learning and higher rewards. Maximizing a mix of reward and information rates–both based on decision confidence, a plausibly accessible quantity to decision-makers during a choice [7, 13, 6]–is a relatively simple and myopic strategy that solves the difficult intertemporal choice problem of how to weigh present rewards against information that can help with future rewards. Previous work has proposed solving this problem via cognitive control, assigning more "effort" to prioritize information over reward when the expectation to learn is high [12], but this strategy, requires a prediction of future discounted reward and is computationally expensive. Our model provides a candidate heuristic that solves this problem naturally and myopically.

A Appendix

A.1 Decision time

Decision time (DT) is defined as reaction time (RT) minus non-decision time t_0. Normalized DT is defined as DT divided by t_0 and RSI.

$$\text{DT} = \text{RT} - t_0 \tag{4}$$

$$\text{normDT} = \frac{\text{DT}}{t_0 + \text{RSI}} \tag{5}$$

A.2 Drift rate & SNR

Drift rate is defined as

$$v = \psi(\mu_A - \mu_B)\frac{dt}{df} \tag{6}$$

where ψ is the stimulus sensitivity, μ_A and μ_B are the means of two normally distributed stimuli A and B, dt is the size of a timestep in the accumulation process (e.g., 1 ms) and df is the stimulus sampling time (e.g., time between frames on a screen, which would be 16 ms at 60 Hz).

The drift rate v increases with increasing ψ, because ψ improves the agent's ability to extract information from the stimuli.

SNR is defined as

$$\text{SNR} = v^2/s_d^2 \tag{7}$$

A.3 Model description detail

An agent accumulates samples according to a standard two-choice drift-diffusion process [14] (Fig.1a). Assuming a Bayesian observer, the accumulator state is converted to a log-posterior ratio of stimulus probability (equation 18 in [6], derived by [13] following [7]).

$$LR_{t=n} = \log\frac{p(S=A|y_{0:n})}{p(S=B|y_{0:n})} = \frac{2v\sum_{t=0}^{n} y}{s_{acc}^2 + ts_d^2v^2} \tag{8}$$

where y is the accumulator, v is the drift rate, s_{acc} is the accumulator noise, and s_d is the drift noise.

Following [6], from eq. 8, we can get the probability of stimulus A given the current evidence

$$p(S=A|y_{0:n}) = \frac{1}{1+e^{LR_{t=n}}} \tag{9}$$

Next, "confidence" $x \in [.5, 1]$ can be computed as the probability of a correct decision assuming a "greedy" response strategy of always choosing the most probable option.

$$x_{0:n} = \max\Big(p(S=A|y_{0:n}), 1 - p(S=A|y_{0:n})\Big) \tag{10}$$

From confidence, the agent can compute $\hat{i}_n$ and $\hat{r}_n$, following eqs. 1, 2 and 3. These quantities are scaled from 0 to 1 separately so that they have comparable values. After scaling, they are combined into $\hat{m}_n$ via a weighted sum where w determines the relative weight of information over reward.

$$\hat{m}_n = w\hat{i}_n^{scaled} + (1-w)\hat{r}_n^{scaled} \tag{11}$$

This mix $\hat{m}_n$ is passed through a Kalman filter for smoothing. A decision is triggered when the smoothed $\hat{m}_n$ first reaches a maximum, i.e. when the smoothed derivative of $\hat{m}_n$ first equals 0. The outcome of the decision is determined by the state of the accumulator at this decision time.

A.4 Learning mechanism

We define a simple across-trial learning model where stimulus sensitivity ψ is an increasing function of cumulative DT.

$$\psi_k = \psi_{max} - (\psi_{max} - \psi_{min})e^{-\alpha \sum_1^n DT_k} \tag{12}$$

where ψ_k is ψ at trial k, ψ_{max} and ψ_{min} are the maximum asymptotic and starting values of ψ respectively, and α is the learning rate.

A.5 Model simulations

For results in Figure 1a-h, we simulated a single trial and visualized model quantities for that single trial.

For results in Figures 1i and 2, we simulated 1000 independent trials with 4000 time steps (nSamples) for each unique parameter combination from the following. If $\hat{m}_n$ was not maximized by the end of the trial, it was assigned a DT $= 4000$ and an outcome corresponding to the accumulator state at that time.

Table 1: Simulation parameters

nSamples	4000	dt	1
s_{acc}	[0.7, 1, 1.3, 1.6]	df	16
s_d	.01	w	[0, 0.25, 0.5, 0.75, 1]
t_0	100	$\mu_A - \mu_B$	1
RSI	[500, 1000, 1500]	ψ	[0.5, 0.75, 1, 1.25, 1.5]

For the iRR-optimal policy, we used the same 1000 accumulator trials and parameter combinations as above (except w because it does not apply), but triggered a choice at the optimal DT for the current SNR and RSI based on the OPC [3, 9]. This produced deterministic DTs but variable ERs, as the outcome of the decision depended on the accumulator state at the optimal DT.

For results in Figure 3, we simulated 1000 sequential trials, where the stimulus sensitivity ψ–and thus the drift rate and SNR–were allowed to change as a function of cumulative DT. Parameters were the same as above, except we fixed $s_d = 1$, RSI $= 1500$, and started with $\psi = 0.1$. Additionally, we set the initial stimulus sensitivity to $\psi_{min} = 0.1$, the asymptotic $\psi_{max} = 2$ and the learning rate $\alpha = 0.00001$, as per the learning mechanism described above.

A.6 DDM fitting

We fit DTs and ERs generated by our model with a standard DDM using the PyDDM Python package [16]. We fit model data with the same parameters described above, choosing intermediate values for the variable parameters $s_d = 1$, $\psi = 0.5$ and RSI $= 1500$. The predicted DT distributions over the model data, as well as the predicted DDM drift and threshold parameters are shown in Figure 1i,j.

To gauge fit quality, DDM data was generated with the fitted parameters and re-fit. The model's BIC was compared to that of 500 simulated DDM datasets of 1000 trials each (matching our model simulation). We calculated a p-value indicating the number of simulation BICs larger (worse) than the model's BIC and found that the model was fit quite well, suggesting it produces plausible DTs and ERs (Fig. 1i insets). For $w = 1$, we found that $p < 0.05$, indicating a poorer fit. Nonetheless, a pure information-rate-maximization policy is relatively less plausible given the considerably long predicted DTs.

We also note that the DDM fits predicted a consistent additional t_0 time of about 200 ms over the DTs. This result is unusual, but we suspect it has to do with the Kalman filter settings which currently prevent very-fast responses. This prediction can be tested and adjusted in future work. Nonetheless, because the model is subject to within-trial assessment of reward and information rates, it will generally predict slower DTs for high noise than a standard DDM where fast decisions can be triggered because of random hits to the threshold.

References

[1] Fuat Balci, Patrick Simen, Ritwik Niyogi, Andrew Saxe, Jessica A Hughes, Philip Holmes, and Jonathan D Cohen. Acquisition of decision making criteria: reward rate ultimately beats accuracy. *Attention, Perception, & Psychophysics*, 73(2):640–657, 2011.

[2] Timothy EJ Behrens, Mark W Woolrich, Mark E Walton, and Matthew FS Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221, 2007.

[3] Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700, 2006.

[4] Rafal Bogacz, Peter T Hu, Philip J Holmes, and Jonathan D Cohen. Do humans produce the speed–accuracy trade-off that maximizes reward rate? *The Quarterly Journal of Experimental Psychology*, 63(5):863–891, 2010.

[5] Scott D Brown and Andrew Heathcote. The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, 57(3):153–178, 2008.

[6] Joshua Calder-Travis, Rafal Bogacz, and Nick Yeung. Bayesian confidence for drift diffusion observers in dynamic stimuli tasks. *BioRxiv*, 2020.

[7] Jan Drugowitsch, Rubén Moreno-Bote, Anne K Churchland, Michael N Shadlen, and Alexandre Pouget. The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, 32(11):3612–3628, 2012.

[8] Joshua I Gold and Michael N Shadlen. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36(2):299–308, 2002.

[9] Philip Holmes and Jonathan D Cohen. Optimality and some of its discontents: Successes and shortcomings of existing models for binary decisions. *Topics in cognitive science*, 6(2):258–278, 2014.

[10] Douglas G Lee and Marius Usher. Value certainty in drift-diffusion models of preferential choice. *Psychological Review*, 2021.

[11] Javier Masís, Travis Chapman, Juliana Y Rhee, David D Cox, and Andrew M Saxe. Rats strategically manage learning during perceptual decision making. *bioRxiv*, 2020.

[12] Javier Alejandro Masís, Sebastian Musslick, and Jonathan Cohen. The value of learning and cognitive control allocation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2021.

[13] Rani Moran. Optimal decision making in heterogeneous and biased environments. *Psychonomic bulletin & review*, 22(1):38–53, 2015.

[14] Roger Ratcliff and Gail McKoon. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922, 2008.

[15] Matthew FS Rushworth and Timothy EJ Behrens. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature neuroscience*, 11(4):389–397, 2008.

[16] Maxwell Shinn, Norman H Lam, and John D Murray. A flexible framework for simulating and fitting generalized drift-diffusion models. *ELife*, 9:e56938, 2020.

[17] Marius Usher and James L McClelland. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 108(3):550, 2001.

[18] Douglas Vickers. Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, 13(1):37–58, 1970.

[19] M Zacksenhouse, R Bogacz, and P Holmes. Robust versus optimal strategies for two-alternative forced choice tasks. *Journal of Mathematical Psychology*, 54(2):230–246, 2010.