Joint Modeling of Choices and Response Times in Multi-stage Decisions via Likelihood Approximation

by

Jialin Li

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Master of Psychology

Department of Psychology

New York University

May, 2026

Professor Marcelo G. Mattar

To my mom, for her endless support and belief in me

Acknowledgements

There are many more people than I would like to acknowledge here. These people include but not limited to the member of my graduate cohort, my graduate advisor, classmates that we've been working on, lifelong friends from my undergraduate and high school.

First and foremost, I would like to express my deepest gratitude to my advisors, Professor Marcelo G. Mattar and Professor Paul W. Glimcher. Thank you for your invaluable guidance, support, and encouragement over the past two years. When I began starting this program, I was at one of the most uncertain stages of my life, questioning whether it was worthwhile to continue pursuing academic research. It was through your encouragement and mentorship that I gradually came to appreciate how intellectually engaging and meaningful this work truly is. Your support has been instrumental in motivating me to continue exploring this field and to pursue the path of becoming a scholar.

Secondly, I would like to thank all the members of the Mattar Lab and the Glimcher Lab for their insightful discussions, valuable suggestions, and continuous support. I am especially grateful to Carlos G. Correa, Fred Callaway, Prakhar Godara, Kenway Louie, Bo Shen, and Duc Nguyen, among others, for their generosity in sharing ideas and providing thoughtful feedback. Their intellectual curiosity and collaborative spirit made this research both productive and enjoyable. I feel particularly fortunate to be part of an environment where challenges can be openly discussed and where colleagues are able to understand complex ideas and offer constructive, insightful responses.

Finally, I would like to thank my family. It is your unconditional support and love that have given me the greatest courage and motivation, enabling me to move forward with determination on the path of academic research. Thank you so, so much.

As I wrote in my undergraduate thesis, each decision carries its own value and meaning, even if it is not immediately apparent. Looking back, I am reminded that no single choice—whether successful or not—defines the trajectory of one's life. Moments of uncertainty and setbacks are not endpoints, but rather part of a broader path that gradually unfolds toward where we are meant to go. Now, this marks a brand-new beginning. Regardless of what lies ahead, I am grateful for the choices I have made, and I move forward with confidence and determination.

This thesis extends the paper accepted at the Proceedings of the 48th Annual Conference of the Cognitive Science Society.

Abstract

Planning involves a process of considering future states before acting. To understand this process, researchers typically infer planning algorithms by fitting computational models to choices. However, different planning models often predict the same choices, despite relying on different computations. Reaction time can help distinguish among models, since different computations produce different temporal signatures. However, incorporating reaction time into fitting is challenging because analytical likelihoods are typically unavailable. Here we propose a likelihood-free method to estimate the density for choices and reaction times in multi-stage decision making. We validate the method through comparisons with analytical solutions, parameter recovery, and showing robust estimates relative to distribution-free and summary statistic approaches. Through a new human experiment and fitting evidence accumulation models from Solway and Botvinick (2015), we demonstrate that modeling the full distribution is important to explain human behavior. Overall, our method is a valuable tool for modeling reaction times in multi-stage decision-making.

Contents

List of Figures

List of Tables

1 Introduction

Every day, people engage in sequential decisions — choosing courses of study, organizing daily routines, or navigating dynamic environments - that require mentally simulating multiple future outcomes before acting. This ability to anticipate and evaluate future consequences is a defining characteristic of planning [Mattar and Lengyel 2022], formalized as a tree search problem in which people construct and search a decision tree that links present actions with distant rewards [Kuperwajs et al. 2025]. Understanding the cognitive and computational mechanisms that enable such planning is therefore essential to explaining how humans make adaptive, goal-directed choices in complex, uncertain environments.

Because planning is an internal, unobservable process, researchers must infer its structure indirectly from behavior. Over the past decade, a rich body of research has modeled human planning by fitting computational models to choices in multi-stage decision-making [Daw et al. 2011]. This approach reveals that people rarely engage in exhaustive planning, instead employing approximate and bounded algorithms that selectively expand promising branches while pruning costly ones [Huys et al. 2012, 2015]. Planning depth and precision adapt flexibly to uncertainty [Fan et al. 2024], working memory capacity [Ying et al. 2024], and effort costs [Callaway et al. 2022], consistent with a resource-rational trade-off between expected reward and cognitive expenditure [Lieder and Griffiths 2020].

While choice data yields numerous insights, additional behavioral measures are often needed for fine-grained algorithmic distinctions. This is because different planning models can produce

the same choice—often the best available option—despite relying on very different computations. To mitigate this model indeterminacy issue, planning models can be compared additionally in their ability to predict human reaction time, since it provides an additional constraint on the underlying computation. Reaction time can, thus, serves as a key behavioral measure for distinguishing among planning algorithms [Callaway et al. 2024; Schwöbel et al. 2024]. Solway and Botvinick [2015] demonstrated this approach by modeling planning dynamics as an evidence integration process. By analyzing reaction time patterns, they compared different forms of evidence integration and found that planning in their task was best described as integrating evidence in parallel across time, with competition occurring across independent paths within the decision tree.

Modeling the temporal dynamics involved in multi-stage decision making, however, poses a methodological challenge. Unlike sequential sampling models (SSMs) used in simple choice tasks [Ratcliff 1978; Brown and Heathcote 2008; Turner et al. 2018], computational models of multi-stage decision making generally lack analytical solutions for the joint probability on choices and reaction times. As a result, researchers have fit models based on summary statistics, or used likelihood-free inference approaches. However, both approaches have important limitations. Fitting summary statistics compresses behavior into a few summary values [Solway and Botvinick 2015], potentially hiding variability in the data. Approximating the likelihood using marginal RT quantiles [Heathcote et al. 2002] assumes independence across stages, an assumption often violated in multi-stage tasks. Inverse binomial sampling (IBS) provides unbiased likelihood estimates for discrete data [van Opheusden et al. 2020], but extending it to continuous variables is nontrivial, leaving detailed aspects of the data distribution unmodeled.

Here we develop a probability density approximation (PDA) method, building on Turner and Sederberg [2014], for estimating likelihoods in multi-stage decision making. We validate PDA using evidence accumulation models from Solway and Botvinick [2015] and a new human planning experiment, demonstrating that PDA produces likelihood estimates closely matching analyt-

ical solutions where available and supports accurate parameter recovery relative to alternative methods. Critically, when applied to human data, fitting the full joint distribution of reaction times—rather than relying on summary statistics or marginal fits—reveals systematic model misspecification undetectable from psychometric curves alone and leads to different model selection conclusions. These results demonstrate that modeling the full reaction time distribution is critical for recovering the temporal dynamics of planning and for reliably evaluating multi-stage decision models.

2 Methods

2.1 Decision tree navigation task

We adapted the sequential decision task from Solway and Botvinick [2015], modifying the spatial layout to prevent participants from using physical proximity as a cue to connectivity. States were arranged in a circle [Callaway et al. 2024; Correa et al. 2023; Zhu et al. 2022], which limits rapid visual scanning for high-reward regions and requires participants to internally track action sequences, thereby increasing reliance on internal planning. The task environment consisted of 11 states, each labeled with a shape indicating the points gained or lost upon visiting (Fig. 2.1A). Arrows indicated available transitions between states. On each trial, participants navigated from an initial state to a terminal state, accumulating points along their trajectory (Fig. 2.1B). To select an action, participants cycled through available options using the 'F' key and confirmed with the 'J' key; decisions were irreversible. Participants' goal was to maximize total points. Trials used one of two tree configurations—'Tree 1' or 'Tree 2'—(Fig. 2.1C), presented in random order.

2.1.1 Experimental procedure

The experiment began with an interactive instruction phase introducing the task structure and rules. Participants then completed guided practice trials, including moving to designated locations to demonstrate task comprehension. To verify learning of the shape–reward associations, participants completed 10 binary-choice practice trials and were required to achieve the maxi-

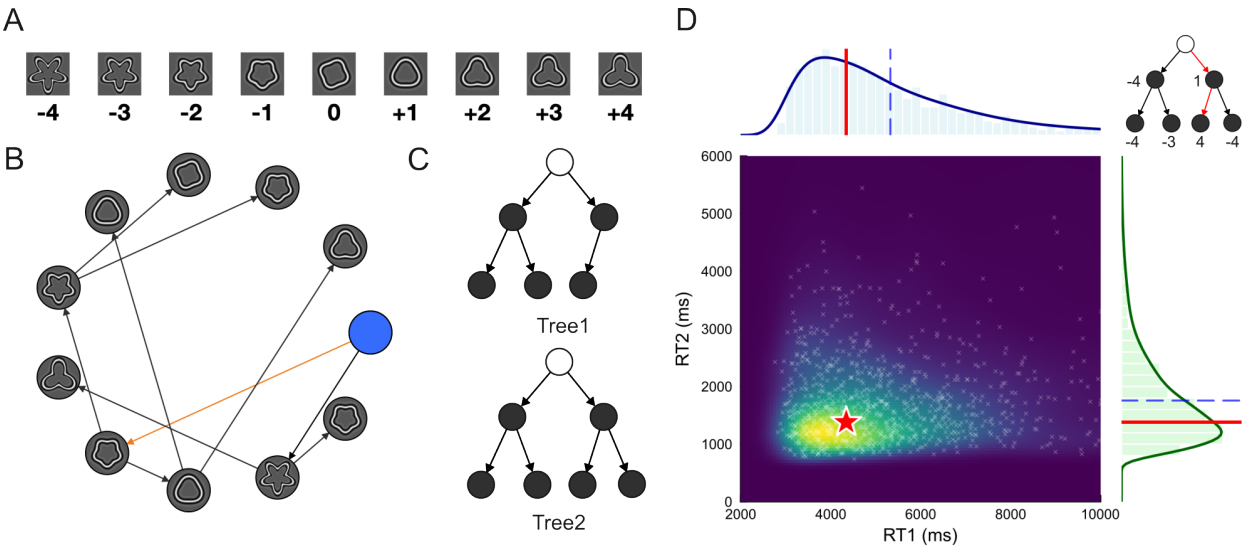

Figure 2.1: Task design and illustration of probability density approximation (PDA) method. (A) The reward associated with each shape ranged from -4 to 4. (B) The task interface shown to the participants. Participants select a node by key press. Their current state is highlighted in blue, and the chosen action is highlighted in orange. The goal is to maximize number of points by navigating in the graph. (C) Two tree configurations. Each node is one state, and each arrow corresponds to an action in the graph. (D) An illustration of PDA. The method draws samples from the simulator model and then uses a two-dimensional Gaussian kernel to estimate the joint probability density of response time. The red lines and star indicate the true reaction time a participant had in this example trial. The blue dashed line indicates the mean of the response time distribution.

mum points on all trials. Participants who did not meet this criterion received additional clarification before proceeding. Following practice, participants completed 200 or 300 experimental trials.

2.1.2 Participants

Forty-five participants were recruited from Prolific who is fluent in English. Following Solway and Botvinick [2015], we excluded trials with response times shorter than 500 ms or longer than 10 s. After exclusions, the final dataset comprised 4772 trials for Tree 1 and 4589 trials for Tree 2. For model fitting, we included only trials with the 'Tree 2' structure, which provides the minimum branching complexity needed to distinguish between the candidate models.

2.2 Statistical Analyses

To examine the effects of first-stage and second-stage accuracy and reaction time across different levels of decision difficulty, we first log-transformed the reaction time data and computed participant-level means. We then conducted repeated-measures ANOVAs separately for accuracy and reaction time. Because each participant completed a limited number of trials and reward configurations were randomly assigned, some difficulty levels in stage 1 contained missing observations. To address this, missing values were imputed using the corresponding population-level mean accuracy and reaction time.

2.3 Computational models

Solway and Botvinick [2015] proposed a family of 13 evidence accumulation models for model-based tree search, differing in their assumptions about noise structure, pruning mechanisms, and search strategies. For brevity, we introduce only the two models used to validate our likelihood-free inference method; full model specifications are provided in Solway and Botvinick [2015].

2.3.1 Forward greedy model (FG)

The forward greedy (FG) model assumes that decisions at each stage are made independently, based only on immediate rewards. At the first stage, evidence for each action accumulates according to:

$$\begin{aligned} E_L^{t+1} &= E_L^t + d_1 \cdot R_L + \epsilon_L, \\ E_R^{t+1} &= E_R^t + d_1 \cdot R_R + \epsilon_R, \end{aligned} \tag{2.1}$$

where E denotes cumulative evidence for a given action, d_1 is the drift rate, R_L and R_R are rewards at the respective next states, and ϵ denotes zero-mean Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ with standard deviation $\sigma = 0.01$. Following Solway and Botvinick [2015], we use L and R to denote the two

actions, though these do not correspond to spatial locations given our circular layout. A decision occurs when the evidence difference exceeds threshold θ_1.

At the second stage, the two states within the chosen branch compete via an analogous process. For example, if action R is selected:

$$\begin{aligned} E_{RL}^{t+1} &= E_{RL}^{t} + d_2 \cdot R_{RL} + \epsilon_{RL}, \\ E_{RR}^{t+1} &= E_{RR}^{t} + d_2 \cdot R_{RR} + \epsilon_{RR}, \end{aligned} \tag{2.2}$$

with threshold θ_2. Each stage includes a non-decision time parameter (T_1 and T_2) accounting for perceptual and motor processes. Integrators reset to zero at each stage.

2.3.2 Two stage independent-path model (TS)

In contrast to the stage-wise FG model, the two-stage independent-path (TS) model treats each complete path through the tree as a single evidence integrator, allowing information about future rewards to influence the first-stage decision. Evidence for each of the four paths is updated according to the cumulative reward along that path:

$$\begin{aligned} E_{L,L}^{t+1} &= E_{L,L}^{t} + (d_1 R_L + \epsilon_L) + (d_1 R_{LL} + \epsilon_{LL}), \\ E_{L,R}^{t+1} &= E_{L,R}^{t} + (d_1 R_L + \epsilon_L) + (d_1 R_{LR} + \epsilon_{LR}), \\ E_{R,L}^{t+1} &= E_{R,L}^{t} + (d_1 R_R + \epsilon_R) + (d_1 R_{RL} + \epsilon_{RL}), \\ E_{R,R}^{t+1} &= E_{R,R}^{t} + (d_1 R_R + \epsilon_R) + (d_1 R_{RR} + \epsilon_{RR}). \end{aligned} \tag{2.3}$$

A first-stage decision occurs when the difference between the best and second-best paths exceeds threshold θ_1. The second stage continues accumulating evidence for the two remaining paths until their difference exceeds θ_2, with $\theta_2 > \theta_1$ to ensure the second-stage decision follows the first. Non-decision time parameters T_1 and T_2 apply as in the FG model.

2.4 Model fitting

We compare three approaches for fitting models to behavioral data: probability density approximation (PDA), which we develop here for multi-stage decisions; residual sum of squares (RSS) [Solway and Botvinick 2015]; and inverse binomial sampling (IBS) [van Opheusden et al. 2020].

2.4.1 Probability density approximation (PDA)

Our approach extends the PDA framework of Turner and Sederberg [2014] to multi-stage decision-making. The key idea is to approximate the likelihood of observed choices and reaction times by simulating the model many times and using kernel density estimation on the resulting samples. PDA requires two components: (1) a dataset $\mathcal{D} = \{\boldsymbol{s}_i, \boldsymbol{c}_i, \boldsymbol{t}_i\}_{i=1}^N$ consisting of N trials, each characterized by a state vector $\boldsymbol{s}_i$, choice vector $\boldsymbol{c}_i$, and reaction time vector $\boldsymbol{t}_i$; and (2) a generative model $\mathcal{M}$ that takes a state $\boldsymbol{s}$ and parameter vector $\boldsymbol{\theta}$ as input and outputs responses $\boldsymbol{r}$ over choices and reaction times.

Assuming independence across trials, the joint likelihood of choices and reaction times given the reward and model parameters can be written as:

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{t}) = \prod_{i=1}^{N} \mathcal{M}(\boldsymbol{c}_i, \boldsymbol{t}_i \mid \boldsymbol{\theta}) \tag{2.4}$$

Following Turner and Sederberg [2014], we factorize the joint distribution using the chain rule:

$$p(c_1, c_2, t_1, t_2 \mid \boldsymbol{\theta}) = p(c_1, c_2 \mid \boldsymbol{\theta})\; p(t_1, t_2 \mid c_1, c_2, \boldsymbol{\theta}) \tag{2.5}$$

We use this identity to compute the empirical likelihood of choices and reaction times in our experimental dataset for each parameter/model combination. The choice probability, $p(c_1, c_2 \mid \boldsymbol{\theta})$,

is estimated from empirical frequencies:

$$p(c_1, c_2 \mid \boldsymbol{\theta}) = \frac{n(c_1, c_2)}{J} \tag{2.6}$$

where $n(c_1, c_2)$ is the number of simulated trials yielding choice pair (c_1, c_2). The conditional RT density, $p(t_1, t_2 \mid c_1, c_2, \boldsymbol{\theta})$, is estimated via multivariate kernel density estimation. Let $\boldsymbol{x} = (t_1, t_2)$ and let $\{\boldsymbol{x}_i\}_{i=1}^{n}$ be the simulated RTs matching the observed choice. The density estimate is:

$$\hat{f}_H(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_H(\boldsymbol{x} - \boldsymbol{x}_i), \tag{2.7}$$

where K_H is a Gaussian kernel with bandwidth matrix H:

$$K_H(\boldsymbol{u}) = \frac{1}{(2\pi)^{d/2}|H|^{1/2}} \exp\left(-\tfrac{1}{2}\boldsymbol{u}^\top H^{-1}\boldsymbol{u}\right) \tag{2.8}$$

The bandwidth matrix H uses a rule-of-thumb scaling based on the sample covariance matrix Σ, which balances bias and variance in the density approximation [Silverman 1986]:

$$H = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \Sigma \tag{2.9}$$

where d is the RT vector dimensionality and Σ is the sample covariance. As $n \to \infty$, $H \to 0$ and the estimate converges to the true density.

Figure 2.1D illustrates PDA for a single trial: the model is simulated repeatedly for a given reward configuration, and the resulting RT samples are used to estimate the joint density (Eq. 2.7). The observed RT (red marker) falls near the density peak, indicating high likelihood under the current parameters.

2.4.2 Residual sum of squares (RSS)

In Solway and Botvinick [2015], for each candidate parameter vector $\boldsymbol{\theta}$, the model was simulated once for every empirical trial, and predictions were summarized into psychometric curves (i.e., choice accuracy and mean RT as functions of task difficulty). The objective function was the residual sum of squares (RSS) between empirical and simulated curves

$$\mathrm{RSS}(\boldsymbol{\theta}) \;=\; \sum_{j=1}^{n} \left(y_j^{\mathrm{data}} - y_j^{\mathrm{sim}}(\boldsymbol{\theta}) \right)^2, \tag{2.10}$$

where y_j^{data} and $y_j^{\mathrm{sim}}(\boldsymbol{\theta})$ are the j-th binned statistics from data and simulation, respectively. Since the accuracy and reaction times are not in the same magnitude, we rescales the RTs. Specifically, the first stage reaction times were divided by 10000 and second-stage reaction times by 1000. This rescaling makes them on more equal weights on the total objective function. Then, BIC values were computed as follows:

$$BIC = k \ln(n) + n \ln(RSS/n) \tag{2.11}$$

where k is the number of parameters, n is the number of data points, and RSS is the residual sum of squares.

2.4.3 Inverse binomial sampling (IBS)

IBS estimates the log-likelihood by repeatedly simulating from the generative model until a simulated response matches the observed one [van Opheusden et al. 2020]. Concretely, for each trial i, simulations are drawn until the first "hit" occurs. The number of simulations required, denoted K_i, follows a geometric distribution with success probability

$$\Pr(K_i = k) = p_i(1 - p_i)^{k-1}, \quad k = 1, 2, \ldots \tag{2.12}$$

An unbiased estimator of the log-likelihood contribution for trial i is then given by

$$\hat{L}_i = \sum_{k=1}^{K_i-1} \frac{1}{k} = \psi(1) - \psi(K_i), \tag{2.13}$$

where $\psi(\cdot)$ is the digamma function. Summing across trials yields the overall IBS estimator:

$$\hat{L}_{\text{IBS}}(\theta) = \sum_{i=1}^{N} \hat{L}_i = \sum_{i=1}^{N} [\psi(1) - \psi(K_i)]. \tag{2.14}$$

For continuous variables such as reaction time, IBS is implemented by introducing a tolerance region around the observed value t_{obs}. That is, a simulated response $t^{(s)}$ is counted as a match if

$$|t^{(s)} - t_{\text{obs}}| \leqslant \delta \tag{2.15}$$

The success probability of this matching process,

$$p_\delta = \Pr\left(|T - t_{\text{obs}}| \leqslant \delta \mid \boldsymbol{\theta}\right). \tag{2.16}$$

provides the basis for the likelihood estimate, where T is the model-generated RT.

2.4.4 Optimization

All three fitting methods require stochastic simulation, making objective function evaluation noisy and computationally expensive. We used Bayesian Adaptive Direct Search (BADS) [Acerbi and Ma 2017; Singh and Acerbi 2024], a derivative-free optimizer designed for expensive, potentially non-smooth objectives. In brief, BADS is a hybrid method that combines Gaussian process-based Bayesian optimization with mesh adaptive direct search (MADS)[Audet and Dennis Jr 2006]. The GP surrogate guides efficient exploration by balancing exploitation uncertainty, while the direct search component provides a robust fallback when the surrogate is inaccurate.

This combination makes it well-suited for our cases, which can offer robust and sample-efficient optimization without requiring gradients.

2.5 Validation of the PDA approach

2.5.1 Comparison with analytical solution

To assess whether PDA recovers accurate likelihood estimates with finite samples, we compared PDA against an analytical solution available for the FG model. Because the FG model treats each stage independently, it can be reparameterized as a drift-diffusion process with tractable first-passage time densities:

$$dx = \mu dt + \sigma dW \tag{2.17}$$

where W is the standard Wiener process and the drift rate $\mu = d_i(R_L - R_R)$ is determined by the reward difference at each stage. Treating each stage as a separate diffusion process, we can analytically compute the first-passage time (FPT) density [Drugowitsch 2016]:

$$\begin{aligned} g_1 &= f_{\text{FPT}}(t_1 \mid \mu_1, \theta_1, \text{upper}_1), \\ g_2 &= f_{\text{FPT}}(t_2 \mid \mu_2, \theta_2, \text{upper}_2), \end{aligned} \tag{2.18}$$

with total log-likelihood:

$$\log p(t_1, t_2 \mid \boldsymbol{c}, \boldsymbol{\theta}) = \log g_1 + \log g_2. \tag{2.19}$$

To compare analytical and PDA likelihood estimates, we first fit the FG model to each participant using the analytical likelihood. We then evaluated PDA log-likelihoods trial-wise at the fitted parameters using $J = 1000$ simulations, requiring at least 100 simulated samples with matching choices.

We found that PDA and analytical log-likelihoods show strong agreement, with high correla-

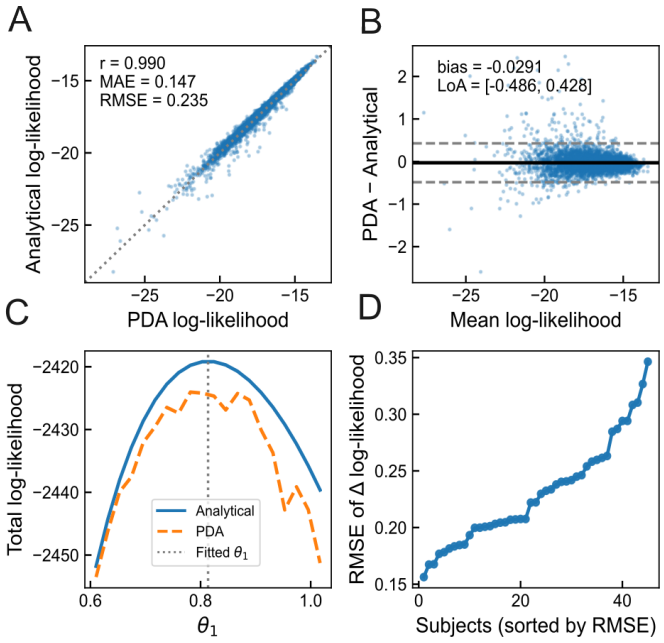

Figure 2.2: Comparison between analytical solution and PDA likelihood estimates in FG model. (A) Trial-wise PDA and analytical log-likelihood show strong correlation. (B) Difference indicates near-zero bias and tight agreement, with no systematic deviation across the likelihood range. (C) A one-parameter likelihood slice around the fitted optimum shows that PDA preserves the local shape and peak of the likelihood landscape. (D) Subject-level RMSE of trial-wise log-likelihood differences remains low across participants.

tion and small, well-controlled differences across the likelihood range (Fig. 2.2A-B). To verify that the PDA estimator preserves not only pointwise likelihood values but also the geometric structure of the likelihood function, for a representative participant, we fixed all model parameters at their fitted values and varied one key parameter (θ_1) over a local range around the optimum. The resulting PDA likelihood curves can preserve the local geometry of the likelihood landscape relevant for parameter optimization (Fig. 2.2C). Finally, subject-level root mean square error (RMSE) are uniformly low, indicating stable agreement across participants (Fig. 2.2D).

2.5.2 Parameter recovery

Parameter recovery assesses whether an inference method can reliably recover true generative parameters under controlled conditions [Wilson and Collins 2019]. We used this approach to

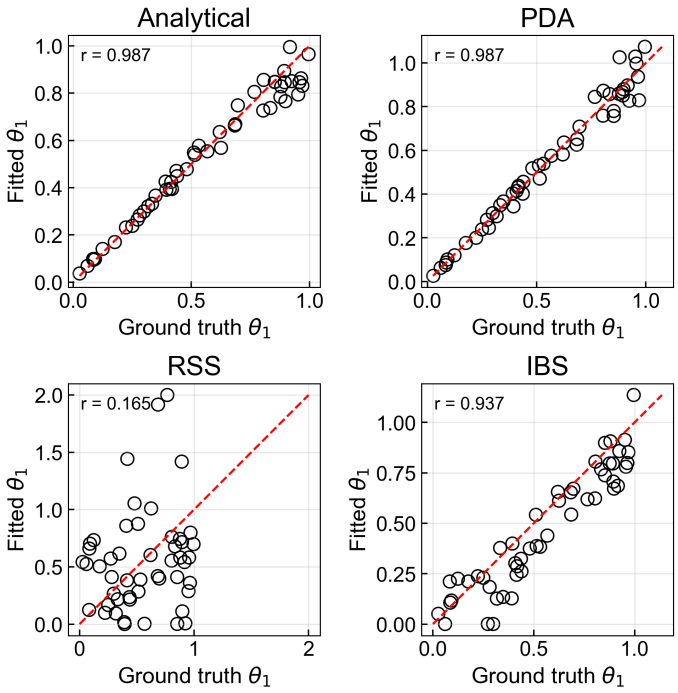

Figure 2.3: Parameter recovery across different fitting method. Analytical likelihood, PDA, RSS and IBS are shown as scatter plots comparing ground truth θ_1 with fitted θ_1 in forward greedy model. Dashed lines indicate the identity line, and r indicates Pearson correlation coefficient.

evaluate PDA relative to RSS and IBS, with analytical likelihood as a benchmark.

We generated 50 synthetic datasets of 100 trials each from the FG model. For each dataset, parameters were sampled from the following distributions: d_1 and d_2 log-uniformly from $[10^{-5}, 10^{-3}]$, θ_1 and θ_2 uniformly from $[0.01, 1]$, and T_1 and T_2 uniformly from $[100, 5000]$ ms. All six parameters were then jointly refit using each method.

Recovery quality differed markedly across methods (Fig. 2.3). PDA achieved near-perfect recovery comparable to the analytical benchmark, with fitted parameters tightly aligned to the identity line. RSS showed substantial variability and systematic deviations, reflecting poor identifiability. IBS performed comparably to PDA in correlation but systematically underestimated decision thresholds, likely because its tolerance-based matching criterion does not fully capture the right-skewed shape of RT distributions.

3 Results

3.1 Human Behavior

We first analyzed whether human choices and response times were systematically modulated by decision difficulty. Consistent with Solway and Botvinick [2015], our repeated-measures ANOVAs showed reliable effects of decision difficulty on both first-stage behavior and second-stage behavior. First-stage choice accuracy varied significantly across difficulty levels ($F(9, 396) = 58.08, p = 5.8810^{-67}, \eta_p^2 = .569$; See Fig.3.1A) [1], indicating that participants were more likely to make the optimal first-stage choice when the best path was more clearly separated from the alternatives. First-stage log reaction time also showed a significant difficulty effect ($F(9, 396) = 49.86, p = 7.6310^{-60}, \eta_p^2 = .531$; See Fig.3.1D), suggesting faster decisions for easier tree-search problems.

At the second stage, accuracy depended significantly on the absolute value difference between the two remaining options ($F(7, 308) = 14.95, p = 8.3610^{-17}, \eta_p^2 = .254$; See Fig.3.1B). Second-stage log reaction time also varied significantly with value difference ($F(7, 308) = 15.83, p = 9.6610^{-18}, \eta_p^2 = .265$; See Fig.3.1E). This pattern mirrors the original finding that second-stage decisions are not merely automatic continuations of a first-stage plan, but instead involve additional value-sensitive deliberation.

[1] These results are computed after imputing the value by population means, However, the results is still significant for only 8 participants who have observations for each difficulty levels.

3.2 Model fitting

In this section, we aimed to replicate the key finding from Solway and Botvinick [2015]—that the TS model best explains human planning behavior. Here, we primarily focus on tree 2 configuration due to its symmetric structure, unlike tree 1 that predict fixed non-decision time if agent choose the branch that only has single path.

Using the original RSS fitting procedure, we simulated all 13 candidate models from that study on our new experimental data, aggregating predictions into group-level psychometric curves and minimizing RSS between empirical and simulated curves. Consistent with the original findings (see Table A.1 in Appendix), the TS model provided the best fit to both accuracy and reaction time psychometric curves at the first stage and second stage (Fig. 3.1[2]). Model comparison using their Bayesian information criterion (BIC) favored the TS model over alternatives (TS: BIC = -289.11; FG: BIC = -257.96).

We next applied our PDA method by fitting model parameters at the individual level to assess the robustness of the results (see Table A.2 in Appendix). In contrast to the previous RSS result, the forward greedy model emerged as the preferred model under individual-level fitting (TS: BIC = 3556.85; FG: BIC = 3523.47). To further examine this discrepancy, we simulated data using individually fitted parameters and constructed group-level psychometric curves. Although both models underestimate choice accuracy at low difficulty levels in both stages relative to human data (Fig. 3.1A and Fig. 3.1B), the most pronounced divergence arises in the second-stage reaction times as a function of second-stage reward difference (Fig. 3.1E), where the TS model differs substantially in its fit.

To elucidate the source of the divergent model comparison results and the qualitative differences observed in the psychometric curves, we draw on insights from Ritz et al. [2026], who demonstrated that apparent model mimicry can arise artefactually from model misspecification.

[2]For brevity, we only plot and discuss the TS and FG models.

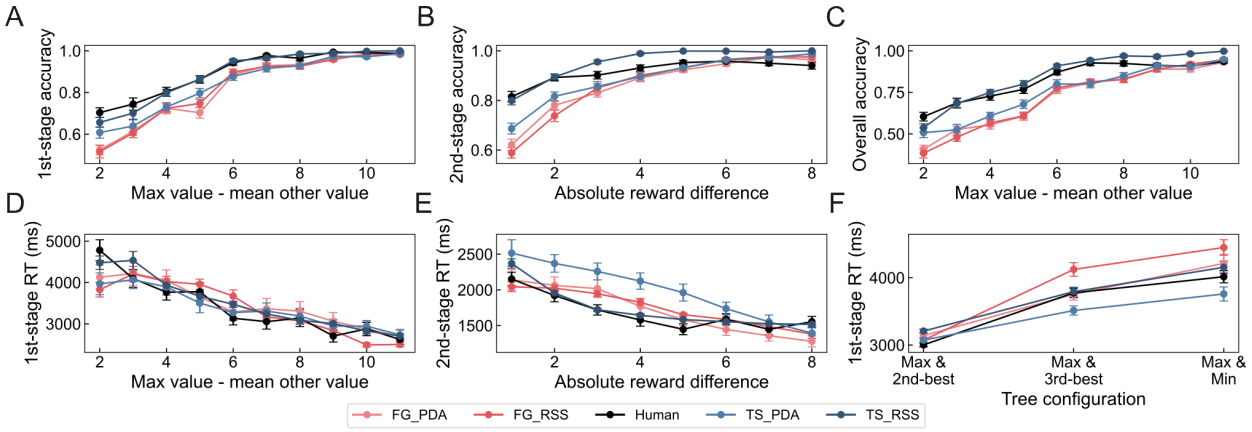

Figure 3.1: Replication of psychometric curves in Solway and Botvinick [2015]. Fits of the TS and FG model using RSS and PDA are shown. (A) First stage choice accuracy as a function of the difference between the maximum path value and the average of other path values. (B) Second stage choice accuracy as a function of the absolute reward difference. (C) Overall accuracy as a function of the difference between the maximum path value and the average of other path values. (D) First stage reaction time versus the difference between the maximum path value and the average of other path values. Only correct trials are included. (E) Second stage reaction time versus absolute reward difference. (F) First stage reaction time across different tree configuration.

Accordingly, we conducted posterior predictive checks using parameters fitted with RSS and PDA, examining the reaction time distributions at both the first and second stages via kernel density estimation. As shown in Fig. 3.2, although the TS model fitted using RSS best reproduces the aggregate psychometric curves, it systematically deviates from the empirical reaction time distributions. This discrepancy is particularly pronounced at the second stage, where the model predicts an unrealistically sharp peak rather than the right-skewed distribution observed in human data. In contrast, while PDA does not fully capture all fine-grained features of the second-stage reaction time pattern (in Fig. 3.1E), it more accurately reproduces the overall distribution (Fig. 3.2). For reference, the reaction time distribution at the second stage is also not well accounted for by the FG model using either RSS or PDA.

Why does the bad prediction at the second stage reaction time distribution only occur in RSS instead of PDA? A plausible explanation related to the assumption in the TS model that evidence continues accumulating between stages. In this model, the first-stage decision threshold is de-

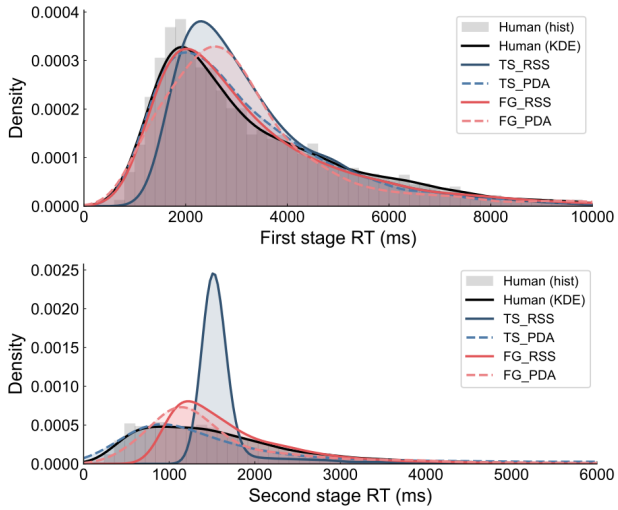

Figure 3.2: Posterior predictive checks of first-stage and second-stage reaction time distributions for the TS and FG models fitted using RSS and PDA. Reaction time distributions are estimated using a Gaussian kernel.

termined solely by the difference between the best and second-best paths. Consequently, upon entering the second stage, the evidence difference within the selected branch may already exceed the second-stage threshold, resulting in near-immediate termination[3] and a small reaction time. This happen in 57.52% of trials when using RSS, but only 24.73% of trials when using PDA. Rapid terminations occur more frequently when second-stage reward differences are large and when the best and second-best paths are in different branches. Because RSS optimizes only mean reaction times and ignores distributional variance, it disproportionately exploits these rapid terminations to account for decreases in second-stage mean reaction time with reward difference. In contrast, PDA penalizes assigning negligible probability mass to regions away from the empirical distribution, preventing such degenerate solutions from maximizing the likelihood and thereby yielding more realistic reaction time distributions. We further validated our hypothesis by fitting models using IBS method and it revealed the similar pattern, in which it favors FG model rather than TS model (TS: BIC = 446.68; FG: BIC = 432.79).

[3]Near-immediate termination is defined as a second-stage decision occurring within 50 ms except non-decision time.

4 Discussion

In this paper, we developed a probability density approximation (PDA) method for fitting cognitive models with continuous variables such as reaction time in multi-stage decision making. Validation against analytical likelihoods and parameter recovery analyses demonstrated that PDA provides robust and reliable estimates, outperforming both summary-statistic (RSS) and tolerance-based (IBS) approaches. Applied to human data, PDA revealed that the model selection conclusion from Solway and Botvinick [2015]—that planning proceeds via parallel integration across independent paths—depends critically on the fitting method: when full RT distributions are modeled rather than summary statistics, a simpler forward greedy model is preferred.

This reversal underscores a broader methodological lesson: good agreement with psychometric curves does not guarantee that a model adequately captures human behavior. In our task, fitting only mean behavior while neglecting distributional shape proved misleading, as model-specific assumptions (specifically, the rapid second-stage terminations in the TS model) artificially improved fits to mean RT and led to incorrect conclusions. PDA facilitates more principled model evaluation by incorporating full distributional information, revealing misspecification that summary statistics conceal. Moreover, individual-level fitting becomes feasible even when the number of trials per participant is limited, enabling examination of individual differences in planning strategies.

Importantly, the improved fit of the FG model under PDA should not be interpreted as evidence that human planning is purely myopic or stage-wise independent. Rather, it suggests that

the specific implementation of the TS model may be misspecificed, particularly in how evidence is accumulated and the propagated across stages. In the TS model, evidence for all complete paths is integrated in parallel, which can lead to excessive accumulation prior to the second stage and consequently near-instantaneous decisions. This mechanism may overestimate the degree to which future outcomes are incorporated during early stages of planning.

From a broader perspective, these findings may also align with resource-rational planning [Callaway et al. 2022], which propose that humans adaptively balance computational cost against expected reward. Instead of performing exhaustive tree search or fully parallel evidence integration, people may rely on simplified or approximate strategies that capture key aspects of the environment while remaining computationally efficient [Kuperwajs et al. 2025]. Integrating behavioral modeling with process-level measures such as eye-tracking may provide further insight into the strategies underlying human planning [Callaway et al. 2024].

In the present study, exploratory analyses of fixation patterns revealed that visual attention is systematically modulated by both uncertainty [Fan et al. 2024] and value [Krajbich et al. 2010; Krajbich and Rangel 2011] (see Appendix). Specifically, participants preferentially allocate attention to nodes that are either more uncertain or more valuable, suggesting a dynamic allocation of cognitive resources during tree search. These findings indicate that existing computational models, such as those assuming fully parallel evidence accumulation [Solway and Botvinick 2015], may not fully capture the mechanisms underlying human planning behavior. Future work should aim to integrate value- and uncertainty-guided attention into computational models, enabling a more fine-grained and process-level account of planning.

We also explore how to derive the optimal policy in the decision tree based on the previous framework [Drugowitsch et al. 2012; Tajima et al. 2016, 2019; Jang et al. 2021] (see Appendix). Under this normative framework based on Bayesian evidence accumulation and dynamic programming, the optimal strategy balances the expected value of committing to a branch against the value of continued information sampling, leading to time-dependent decision boundaries that

collapse as the cost of deliberation increases. Although there are several critical assumptions in our framework, incorporating these principles from the optimal policy-such as uncertainty-sensitive stopping rules and belief propagation-may provide a promising direction for developing more accurate and mechanistically grounded models of multi-stage planning. Future work may also consider combining with neural network method to approximate the optimal policy when the derivation of such planning task is intractable [Chen et al. 2026].

Future work could extend the present framework in several important ways. First, the accuracy of PDA likelihood estimates depends critically on the number of simulated samples used for kernel density estimation. Indeed, small sample sizes can introduce substantial bias, particularly for low-probability events. Second, neither model we tested fully captured human RT distributions, suggesting that more flexible model architectures may be needed. Future work could address sample efficiency by incorporating importance sampling [Tran et al. 2020] or by using neural networks to approximate likelihood functions [Fengler et al. 2021]. In addition, prior work has shown that omission trials can substantially affect parameter estimates [Leng et al. 2024]; integrating omissions into our framework represents another promising direction.

A | Appendix

A.1 Optimal policy in decision tree

A.1.1 Generative Model

We consider a sequential evidence accumulation framework in which the latent reward of each option (or node in the decision tree) is denoted by z_i. The agent does not observe z_i directly, but instead receives noisy samples over time.

Mathematically, At each time step t, the agent receives a noisy and momentary observation increment from the latent variable:

$$\delta x_i(t) \sim \mathcal{N}(z_i\, \delta t, \sigma_x^2\, \delta t) \tag{A.1}$$

then the accumulated evidence evolves as:

$$x_i(t) = \sum_{n=1}^{N} \delta x_i(n), \quad t = N\delta t \tag{A.2}$$

We assume a Gaussian prior over latent rewards:

$$z \sim \mathcal{N}(\bar{z}, \Sigma_z) \tag{A.3}$$

A.1.2 Optimal Evidence Accumulation

Here, we define optimal accumulation process in the sense that it integrates evidence over time while accounting for uncertainty reduction. Under Gaussian assumptions, the posterior remains Gaussian:

$$p(z \mid x(t)) \propto p(x(t) \mid z)\, p(z) \tag{A.4}$$

with posterior mean and covariance:

$$\mu(t) = \Sigma(t)\left(\Sigma_z^{-1}\bar{z} + \Sigma_x^{-1}x(t)\right), \quad \Sigma(t) = \left(\Sigma_z^{-1} + t\Sigma_x^{-1}\right)^{-1} \tag{A.5}$$

That is, the agent sequentially updates beliefs using Bayes' rule and posterior variance decreases with time.

A.1.3 Optimal Stopping Rule

The optimal policy is derived using dynamic programming, Let $V(z,t)$ denote the value function, which represents the maximal expected return given current belief state z at time t. The Bellman equation is:

$$V(z,t) = \max\left\{Q_{\text{wait}}(z,t), Q_{\text{left}}(z), Q_{\text{right}}(z)\right\} \tag{A.6}$$

The value of waiting is given by the expected future value minus a time cost:

$$Q_{\text{wait}}(z,t) = \int V(z', t+\Delta t)\, K(z' \mid z)\, dz' - c \cdot \Delta t \tag{A.7}$$

where $K(z' \mid z)$ is the transition kernel induced by Bayesian belief updating, c is the cost of time, Δt is the time step. Different from simple choice decision making [Tajima et al. 2016, 2019] where the observation is independent across different items, in our decision tree setup, the path

value is correlated since two paths within the same branch share the same root node. To take this correlation into account, we assume that the belief state is two-dimensional and evolves as:

$$K(z_1', z_2' \mid z_1, z_2) = \mathcal{N}(z_1' \mid z_1, \sigma_1^2)\, \mathcal{N}(z_2' \mid z_2, \sigma_2^2) \tag{A.8}$$

where

$$(z_1, z_2) \sim \mathcal{N}(\mu, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{A.9}$$

To examine how correlation shapes decision-making under this framework, we consider a simplified setting with two options and varying levels of correlation ($\rho = -0.5, 0, 0.5$). Correlation influences the relative value of further sampling by altering the joint uncertainty structure, thereby reshaping the geometry of the decision boundaries. Consistent with prior work, the decision boundaries collapse over time, reflecting an increasing urgency to commit as the cost of sampling accumulates [Tajima et al. 2016, 2019]. Notably, we find that higher correlation reduces the magnitude of the decision boundary, indicating that correlated evidence facilitates faster decisions by diminishing the benefit of additional information sampling.

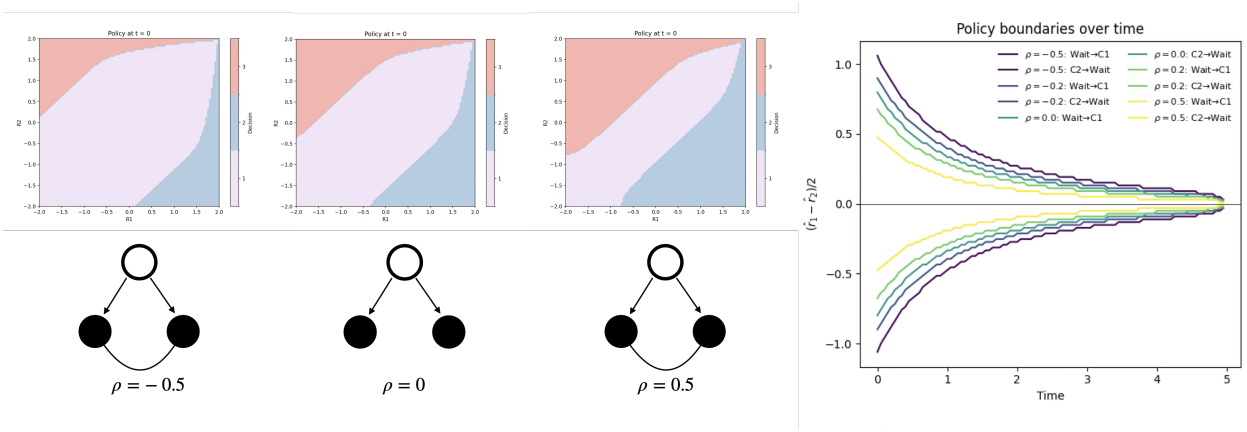

Figure A.1: Effect of correlation on optimal decision boundaries. Left: Policy regions at $t = 0$ for different correlation levels ($\rho = -0.5, 0, 0.5$). Right: Decision boundaries collapse over time, reflecting increasing urgency. Higher correlation reduces boundary magnitude, leading to faster decisions. Higher correlation leads to smaller boundary magnitudes, indicating reduced value of further sampling and faster commitment under correlated evidence.

For choice actions, the agent evaluates the expected reward of that subtree when committing to a subtree. Here, we define its value by the maximum over its terminal nodes:

$$Q_{\text{left}}(z) = \mathbb{E}[\max(z_{LL}, z_{LR})] \quad Q_{\text{right}}(z) = \mathbb{E}[\max(z_{RL}, z_{RR})] \tag{A.10}$$

Thus, the agent compares the expected values of that subtree values rather than individual nodes. To compute this value, we use the Clark formula [Clark 1961]:

$$\mathbb{E}[\max(X, Y)] = \mu_1 \Phi(\delta) + \mu_2 \Phi(-\delta) + \theta\, \phi(\delta) \tag{A.11}$$

where

$$\theta = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}, \quad \delta = \frac{\mu_1 - \mu_2}{\theta} \tag{A.12}$$

and Φ, ϕ are the standard normal CDF and PDF. This provides an analytic solution for subtree valuation under uncertainty and correlation.

Figure A.2 illustrates how the expected value of a subtree, defined as $\mathbb{E}[\max(X, Y)]$, varies with the posterior means of two variables. In tree search, this corresponds to evaluating a branch by taking the maximum over its path nodes. The surface shows that the expected value increases monotonically with both means, with a smooth nonlinear transition near the diagonal where the two values are similar. As reflected in the contour plot, the value function behaves as a soft maximum, rather than a purely additive or winner-take-all rule.

A.1.4 Optimal Policy Structure

To characterize the structure of the optimal policy, we solved the Bellman equation over a discretized belief space using dynamic programming. As the optimal policy is defined in the 5-dimensional space (4 belief dimensions + 1 temporal dimension), we ran a toy version to show its overall policy structure. Specifically, the belief space was defined over a grid of posterior means,

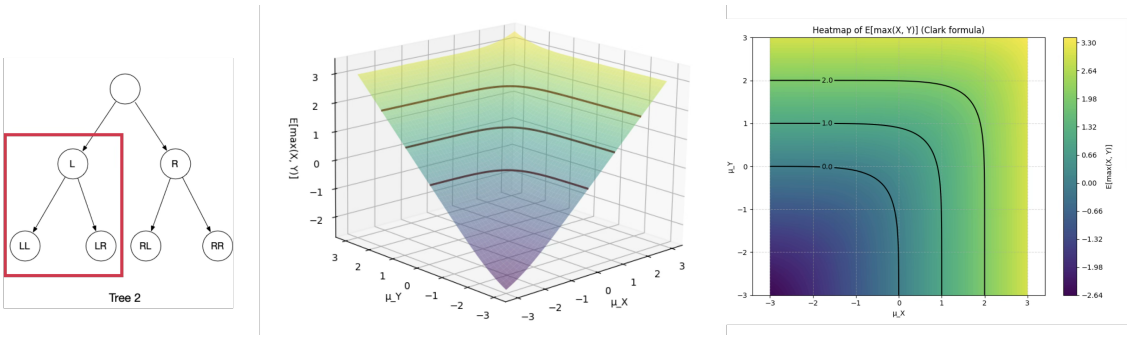

Figure A.2: Expected subtree value as a function of posterior means. Left: Tree structure highlighting a subtree (LL, LR) whose value is computed as $\mathbb{E}[\max(X, Y)]$. Middle: Surface plot of $\mathbb{E}[\max(X, Y)]$ as a function of posterior means μ_X and μ_Y. Right: Corresponding contour plot. The smooth curvature near the diagonal reflects uncertainty integration, where both options contribute to the expected value when their means are similar.

with 50 belief grid points per dimension and 100 time steps, and truncated within a bounded range of $z_{\max} = 3$. The temporal resolution was set to $\Delta t = 0.1$, and a constant time cost $c = 0.5$.

Due to its high dimensionality of policy structure, we can only plot it in the 2-dimensional space once we fixed the other two path values. Figure A.3 illustrates how the optimal policy is modulated by the values of different subtrees in the decision tree. The top row shows the policy as a function of the values of the left subtree (LL, LR), while holding the right subtree fixed. As the values of the left subtree increase, the decision region corresponding to choosing the left branch expands, while the waiting region shrinks. This indicates that higher expected subtree value reduces the need for further sampling and promotes earlier commitment.

The bottom row instead varies the values along the LR–RL paths, while keeping the remaining nodes (e.g., LL and RR) fixed. In this case, the decision boundary is shaped by the relative value between competing paths across the two subtrees. As the value difference between LR and RL increases, the policy increasingly favors the higher-value branch, leading to a shift from waiting to immediate commitment. When the two paths have similar values, a broad waiting region emerges, reflecting increased uncertainty.

Several critical assumptions underlie the derivation of the optimal policy presented here. First, the application of the Clark formula is restricted to the case of two (possibly correlated) Gaussian

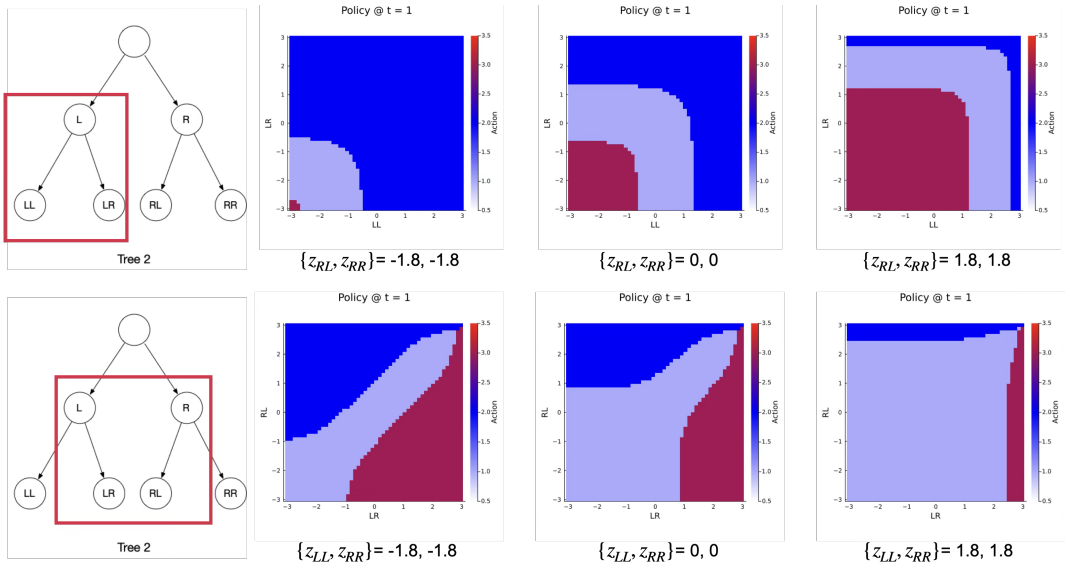

Figure A.3: Effect of subtree and path values on the optimal policy. Top row: Policy as a function of left subtree values (LL, LR), with the right subtree fixed. Bottom row: Policy as a function of path values along LR and RL, while holding other nodes fixed. Colors indicate actions (choose left, choose right, or wait). Increasing the value of a subtree expands its corresponding decision region. When competing paths have similar values, a larger waiting region emerges, reflecting increased uncertainty and the value of further sampling.

variables [Clark 1961]. When extending the breadth of the subtree to include more than two terminal nodes, no closed-form analytical solution exists for computing the expected maximum, and approximation methods—such as simulation-based techniques—are required.

Second, the optimal policy derived in this framework applies primarily to the first stage of decision-making, in which the agent accumulates evidence and selects between two branches. In the experimental task, however, participants subsequently make a second-stage decision after committing to a branch. Although this second-stage decision can be approximated by standard sequential sampling models [Tajima et al. 2016, 2019], it is important to note that the evidence accumulated during the first stage should, in principle, be retained and carried forward. The current formulation does not explicitly model this propagation of belief across stages. Future work could address this limitation by incorporating mechanisms to preserve and update beliefs across hierarchical decisions, for example through extensions based on nested dynamic programming.

A.2 Eye-tracking results

In addition to the behavioral results shown in the main texts, we also collected 17 participants' eye-tracking data during the task, in order to better understand the evidence integration and strategy underlying decision-making. Fixation patterns provide a window into how participants allocate attention across nodes in the decision tree [Callaway et al. 2024], and whether such allocation is consistent with value-based or uncertainty-based strategies [Fan et al. 2024].

First, we categorized fixations according to node depth within the tree. Across both tree configurations, participants exhibited a strong bias toward first-layer nodes during the initial stage, with significantly longer fixation durations compared to second-layer nodes. After committing to a branch, attention shifted toward second-layer nodes to further refine the decision. This pattern suggests that participants do not allocate cognitive resources uniformly across all nodes, but instead adapt their attention based on the hierarchical structure of the task. Such behavior is consistent with a resource-rational strategy, in which limited cognitive resources are selectively deployed to maximize decision-relevant information [Lieder and Griffiths 2020]

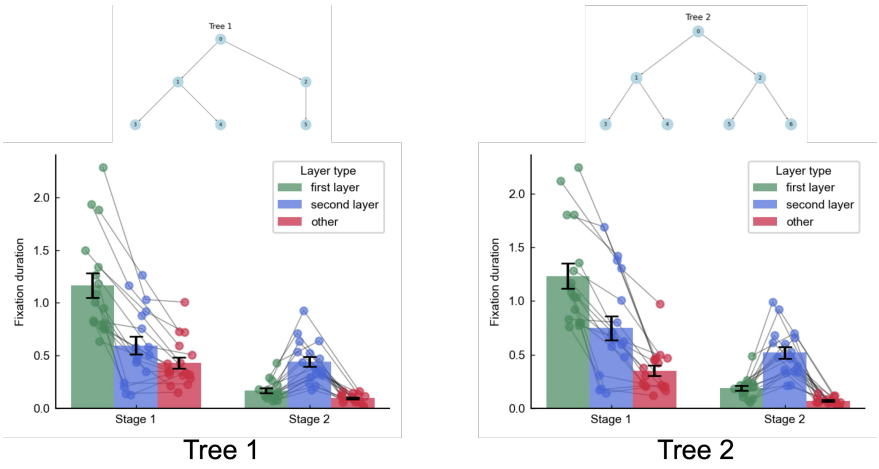

Figure A.4: Fixation duration across node layers and decision stages for Tree 1 and Tree 2. Fixations are categorized into first-layer nodes, second-layer nodes, and other regions. During Stage 1, participants exhibit longer fixation durations on first-layer nodes, indicating an initial focus on high-level branch evaluation. In Stage 2, fixation shifts toward second-layer nodes, reflecting refinement within the selected subtree. Error bars denote ± s.e.m., and dots represent individual participants.

To examine how reward influences fixation behavior, we further categorized fixations based on the absolute reward difference between the two first-layer nodes. Fixation duration systematically decreased as the reward difference increased: when the two options were similar in value, participants spent more time fixating, whereas larger differences led to shorter fixation durations. This pattern is consistent with uncertainty-driven information sampling [Fan et al. 2024], whereby small reward differences correspond to high decision uncertainty and thus increased sampling, while large differences reduce uncertainty and facilitate faster decisions.

In addition, fixation duration was consistently longer for chosen nodes compared to unchosen nodes, reflecting a value-driven allocation of attention. This finding aligns with previous work demonstrating that visual attention is biased toward higher-value options and plays an active role in shaping choice behavior [Krajbich et al. 2010; Krajbich and Rangel 2011].

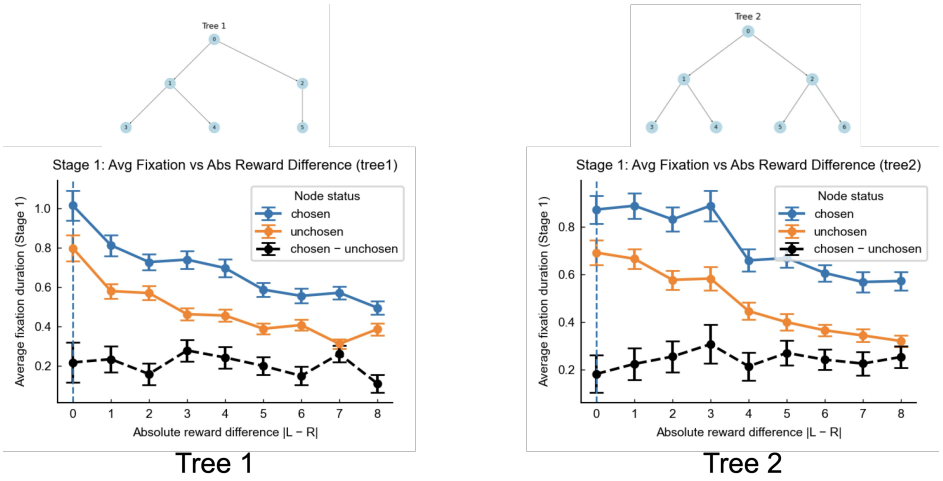

Figure A.5: Relationship between fixation duration and absolute reward difference between first-layer nodes ($|L - R|$) during Stage 1 for both tree configurations. Fixation duration decreases as reward difference increases, indicating reduced sampling under lower uncertainty. Blue and orange lines denote chosen and unchosen nodes, respectively, while the dashed black line shows their difference. Error bars denote ± s.e.m. This pattern is consistent with uncertainty-driven information sampling.

We further examined fixation duration across different tree configurations, categorized by the relative ranking of path values. Fixation patterns varied systematically as a function of path value, with higher-value paths receiving longer fixation durations. This result suggests that attention is

not merely stimulus-driven, but is guided by value. Specifically, participants selectively allocate attention to nodes that are more likely to influence the final decision, consistent with a value-directed information sampling strategy.

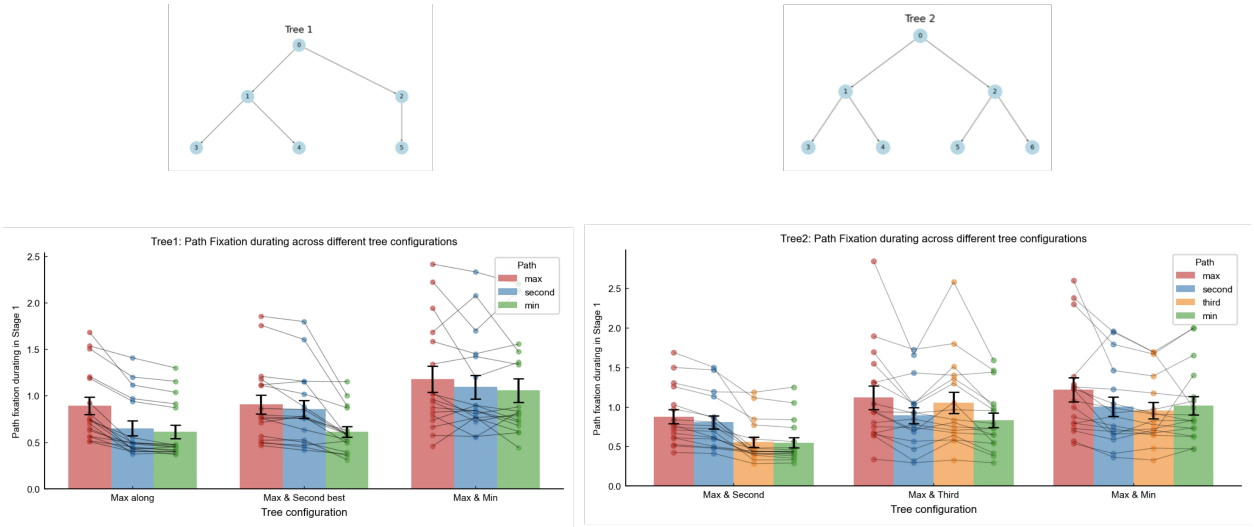

Figure A.6: Fixation duration allocated to different paths across tree configurations. Each configuration manipulates the relative ranking of path values. Participants allocate more fixation time to higher-value paths, with the strongest bias toward the maximum-value path. Lines connect individual participants, and bars show group means with ± s.e.m., indicating consistent value-directed attention across configurations.

Finally, we examined the relationship between fixation duration and choice probability at the first stage. The probability of selecting an option increases monotonically with the amount of fixation it receives, an effect that is strongest for high-value nodes and consistent across both tree structures. Notably, the influence of fixation duration on choice is most pronounced for nodes with intermediate reward values, where decision uncertainty is highest. These results provide further evidence that value-directed fixation plays a causal role in shaping choice. Within the framework of the optimal policy [Callaway et al. 2021; Jang et al. 2021], fixation can be interpreted as a form of selective information sampling, whereby the agent preferentially gathers evidence from the most informative nodes.

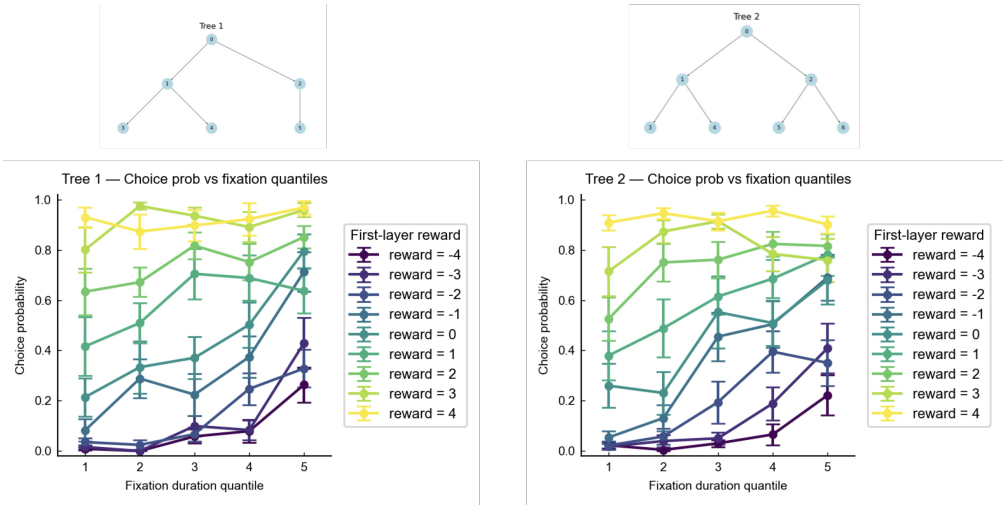

Figure A.7: Choice probability as a function of fixation duration quantiles for different first-layer reward levels in Tree 1 and Tree 2. Higher fixation duration is associated with increased probability of choosing the corresponding option, with stronger effects for higher-value options. Error bars denote ± s.e.m. These results indicate a systematic relationship between attention allocation and decision outcomes, consistent with value-directed evidence accumulation.

A.3 Model Comparison

The table summarizes model fits across RSS, PDA, and IBS, with BIC used for comparison. Consistent with prior work, the two-stage independent paths model provides the best fit under RSS. However, likelihood-based methods (PDA and IBS) yield a different pattern, with the forward greedy model achieving the lowest BIC, indicating a better fit under these frameworks.

We did not include the one-stage vigor model and its variants from PDA fitting. These models assume second-stage reaction times are driven solely by reward-dependent vigor and fail to capture continued deliberation, leading to systematic underestimation of second-stage accuracy. Consistent with prior findings [Solway and Botvinick 2015], our preliminary IBS results also showed inferior performance, so we did not further evaluate these models under PDA.

Table A.1: Model comparison based on RSS method

Models	Types	Parameters	Num of Pars	RSS
1	Two-stage, independent paths (primary model)	d1, d2, θ_1, θ_2, T1, T2	6	**-289.11**
2	Two-stage, correlated paths	d1, d2, θ_1, θ_2, T1, T2	6	-280.81
3	Two-stage, independent paths, with pruning	d1, d2, θ_1, θ_2, T1, T2, θ_{prun}	7	-279.50
4	Two-stage, correlated paths, with pruning	d1, d2, θ_1, θ_2, T1, T2, θ_{prun}	7	-276.48
5	Two-stage, independent paths, single drift	d, θ_1, θ_2, T1, T2	5	-281.45
6	Forward greedy	d1, d2, θ_1, θ_2, T1, T2	6	-257.96
7	Backward search	d0, d1, d2, θ_0, θ_1, θ_2, T1, T2	8	-265.44
8	Backward search, same first-stage parameters	d0, d2, θ_0, θ_2, T1, T2	6	-254.66
9	Backward search with reset	d0, d1, d2, θ_0, θ_1, θ_2, T1, T2	8	-273.89
10	Backward search with reset, same first-stage	d0, d2, θ_0, θ_2, T1, T2	6	-278.17
11	One-stage with vigor	d, θ, vigor1, vigor2, T1, T2	6	-
12	One-stage with single vigor	d, θ, vigor, T1, T2	5	-
13	One-stage with single vigor and rating noise	d, θ, vigor1, vigor2, T1, T2, rating_sd	7	-

Table A.2: Model comparison based on PDA method

Models	Types	Parameters	Num of Pars	PDA
1	Two-stage, independent paths (primary model)	d1, d2, θ_1, θ_2, T1, T2	6	3556.85
2	Two-stage, correlated paths	d1, d2, θ_1, θ_2, T1, T2	6	3556.42
3	Two-stage, independent paths, with pruning	d1, d2, θ_1, θ_2, T1, T2, θ_{prun}	7	3576.54
4	Two-stage, correlated paths, with pruning	d1, d2, θ_1, θ_2, T1, T2, θ_{prun}	7	3564.25
5	Two-stage, independent paths, single drift	d, θ_1, θ_2, T1, T2	5	3555.47
6	Forward greedy	d1, d2, θ_1, θ_2, T1, T2	6	**3523.46**
7	Backward search	d0, d1, d2, θ_0, θ_1, θ_2, T1, T2	8	3642.90
8	Backward search, same first-stage parameters	d0, d2, θ_0, θ_2, T1, T2	6	3645.45
9	Backward search with reset	d0, d1, d2, θ_0, θ_1, θ_2, T1, T2	8	3650.93
10	Backward search with reset, same first-stage	d0, d2, θ_0, θ_2, T1, T2	6	3644.01
11	One-stage with vigor	d, θ, vigor1, vigor2, T1, T2	6	-
12	One-stage with single vigor	d, θ, vigor, T1, T2	5	-
13	One-stage with single vigor and rating noise	d, θ, vigor1, vigor2, T1, T2, rating_sd	7	-

Table A.3: Model comparison based on IBS method

Models	Types	Parameters	Num of Pars	IBS
1	Two-stage, independent paths (primary model)	d1, d2, θ_1, θ_2, T1, T2	6	446.68
2	Two-stage, correlated paths	d1, d2, θ_1, θ_2, T1, T2	6	472.17
3	Two-stage, independent paths, with pruning	d1, d2, θ_1, θ_2, T1, T2, θ_{prun}	7	442.89
4	Two-stage, correlated paths, with pruning	d1, d2, θ_1, θ_2, T1, T2, θ_{prun}	7	445.64
5	Two-stage, independent paths, single drift	d, θ_1, θ_2, T1, T2	5	464.62
6	Forward greedy	d1, d2, θ_1, θ_2, T1, T2	6	**432.79**
7	Backward search	d0, d1, d2, θ_0, θ_1, θ_2, T1, T2	8	447.86
8	Backward search, same first-stage parameters	d0, d2, θ_0, θ_2, T1, T2	6	455.92
9	Backward search with reset	d0, d1, d2, θ_0, θ_1, θ_2, T1, T2	8	449.22
10	Backward search with reset, same first-stage	d0, d2, θ_0, θ_2, T1, T2	6	451.75
11	One-stage with vigor	d, θ, vigor1, vigor2, T1, T2	6	530.06
12	One-stage with single vigor	d, θ, vigor, T1, T2	5	532.74
13	One-stage with single vigor and rating noise	d, θ, vigor1, vigor2, T1, T2, rating_sd	7	530.18

Bibliography

Acerbi, L. and Ma, W. J. (2017). Practical bayesian optimization for model fitting with bayesian adaptive direct search.

Audet, C. and Dennis Jr, J. E. (2006). Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on optimization*, 17(1):188–217.

Brown, S. D. and Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57(3):153–178.

Callaway, F., Rangel, A., and Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology*, 17(3):e1008863.

Callaway, F., Van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., and Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, 6(8):1112–1125.

Callaway, F., Yu, M., and Mattar, M. G. (2024). Revealing human planning strategies with eye-tracking. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Chen, S., Callaway, F., Kumar, S., Lupkin, S. M., Wallis, J. D., McGinty, V. B., Rich, E. L., and Mattar, M. G. (2026). Learning to select computations in recurrent neural circuits. *bioRxiv*, pages 2026–04.

Clark, C. E. (1961). The greatest of a finite set of random variables. *Operations Research*, 9(2):145–162.

Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., and Griffiths, T. L. (2023). Humans decompose tasks by trading off utility and computational cost. *PLoS Computational Biology*, 19(6):e1011087.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215.

Drugowitsch, J. (2016). Fast and accurate monte carlo sampling of first-passage times from wiener diffusion models. *Scientific Reports*, 6(1):20490.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, 32(11):3612–3628.

Fan, H., Callaway, F., and Gershman, S. J. (2024). Uncertainty-driven exploration during planning.

Fengler, A., Govindarajan, L. N., Chen, T., and Frank, M. J. (2021). Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience. *eLife*, 10:e65074.

Heathcote, A., Brown, S., and Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, 9(2):394–401.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J. P. (2012). Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees. *PLoS Computational Biology*, 8(3):e1002410.

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., and Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, 112(10):3098–3103.

Jang, A. I., Sharma, R., and Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *Elife*, 10:e63436.

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, 13(10):1292–1298.

Krajbich, I. and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857.

Kuperwajs, I., Russek, E. M., Mattar, M. G., Ma, W. J., and Griffiths, T. L. (2025). Looking deeper into the algorithms underlying human planning. *Trends in Cognitive Sciences*, 0(0).

Leng, X., Fengler, A., Shenhav, A., and Frank, M. J. (2024). The Perils of Omitting Omissions when Modeling Evidence Accumulation. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)*.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1.

Mattar, M. G. and Lengyel, M. (2022). Planning in the brain. *Neuron*, 110(6):914–934.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2):59–108.

Ritz, H., Frömer, R., and Shenhav, A. (2026). Misspecified models create the appearance of adaptive control during value-based choice. *Communications Psychology*, 4(1):11.

Schwöbel, S., Marković, D., Smolka, M. N., and Kiebel, S. (2024). Joint modeling of choices and reaction times based on bayesian contextual behavioral control. *PLOS Computational Biology*, 20(7):e1012228.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Routledge, London: Chapman & Hall.

Singh, G. S. and Acerbi, L. (2024). PyBADS: Fast and robust black-box optimization in python. *Journal of Open Source Software*, 9(94):5694.

Solway, A. and Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, 112(37):11708–11713.

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

Tajima, S., Drugowitsch, J., and Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature communications*, 7(1):12400.

Tran, M.-N., Scharth, M., Gunawan, D., Kohn, R., Brown, S. D., and Hawkins, G. E. (2020). Robustly estimating the marginal likelihood for cognitive models via importance sampling. *Behavior Research Methods*, 53(3):1148–1165.

Turner, B. M., Schley, D. R., Muller, C., and Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, 125(3):329–362.

Turner, B. M. and Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, 21(2):227–250.

van Opheusden, B., Acerbi, L., and Ma, W. J. (2020). Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *PLOS Computational Biology*, 16(12):e1008483.

Wilson, R. C. and Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, 8:e49547.

Ying, Z., Callaway, F., Kiyonaga, A., and Mattar, M. G. (2024). Resource-rational encoding of reward information in planning. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)*.

Zhu, S., Lakshminarasimhan, K. J., Arfaei, N., and Angelaki, D. E. (2022). Eye movements reveal spatiotemporal dynamics of visually-informed planning in navigation. *Elife*, 11.